Biostatistical and meta-research approaches to assess diagnostic test use



Jack William O'Sullivan

Green Templeton College

University of Oxford

A thesis submitted for the degree of *Doctor of Philosophy*

Trinity term, 2018

Acknowledgements

Rarely does one get an opportunity to thank those that have propped us up, to directly express gratitude and love. The acknowledgements section of a DPhil, to me, is one of these rare opportunities; one that must be embraced.

When I first arrived in Oxford, my supervisors handed me their former student's DPhil theses - "to help you get a feel for what is expected". Admittedly, I didn't read much beyond my predecessor's acknowledgements page. Their reflections intrigued me more than the caveats of their methodological approach. All that I read started by thanking their supervisors. This seemed logical at the time but only now, at the conclusion of my DPhil, do I understand why this is the norm. Put simply, albeit unoriginally, I would not have produced this body of work without my supervisor's support and guidance.

I am grateful to Professor Carl Heneghan for many things, but particularly his creative approach to research and his mentorship. Carl, you have shown me how to generate research questions that matter and how to tackle problems from novel angles. Your mentorship has taught me how to make things happen.

I have benefited enormously from Professor Rafael Perera's expertise and guidance. Raf has a unique gift; he understands the intricacies of a research project, and the bigger picture – the study's overall goal. I've learnt a lot about statistics and I am immensely grateful for his professional and personal support over the years.

Dr Jeffrey Aronson can take credit for transforming my writing from terrible to (hopefully) readable. Admittedly, I suspect Jeff still believes this transformation has yet to occur. I have many fond memories of Jeff, but one that particularly stands out involves (arguably) England's most eminent writer. To explain the grammatical difference between "compare to" and "compare with", Jeff recited, by heart, William Shakespeare's Sonnet *Shall I compare thee to a summer's day*. T'was wonderful.

Although they didn't indulge me with the intricacies of the English language, I am immensely grateful to my parents. My work as a doctor has exposed me to harsh reality of many people's lives. More people than you'd care to imagine have no one visit them, no one care for them. This insight into an array of people's lives has made me appreciate (even more) the unwavering love and support I received, and continue to receive, from my parents. It is comforting to know, no matter what, they will be there for me. Seemingly trivial conversations with my parents often help me feel grounded, and always make me feel loved.

My brothers have been a source of inspiration during this thesis. Although I often feel I missed your transformation from boys to men, I really enjoy the time we get to spend together now. I'd love to see more of you, although we currently seem destined to biannual (very enjoyable) family holidays. Perhaps my move closer to Harry's adopted home land will prompt a transpacific move for Mum, Dad and Tom. Regardless, you guys mean a lot and, when I speak of you, it is with real pride and love.

My family know me well, but unfortunately (or fortunately) I feel I have undergone many years of personal growth beyond their direct vision. There are two groups of people who have witnessed this maturation directly, and they deserve (lots of) recognition. First, my Oxford family. In no particular order, I would like to thank the following people for riding the highs and the lows, offering advice and looking out for my well-being. Sven Jaeschke: for always being up for a good time and a good chat, Tori Pereira: for your energy, enthusiasm and reflections (and for always matching my energy levels), Andrew Linden: for your steadfast approach to life and all the fun times (both with and without spirits), Jason Hendry: for your curiosity both professionally and personally and for recalibrating my vernacular (hooking… will never have the same meaning for me), Ali Albasri: for your kind and sensible words – they always dissipated any emotional or professional tumult, Dylan James: for being as close to a rock as one can be, Ellie Price for all the laughter and good times, and Alexander Finlayson: for your important insights and thoughts. I was also fortunate enough to befriend Mark Glaire and William Warr late in my DPhil, who I also count as close friends.

I suspect all my Oxford friends would unequivocally agree that I have changed over the last three years. It is slightly worrying to reflect on the change my Australian friends witnessed during my time at medical school/junior doctoring (age 18 to 24). To Brad Strange, Rob Harvey, Greg Leeb, Fadi Hosn, Lachy Cooper, Josh Beaver, and Michael Lloyd; you guys mean a lot and I am very thankful for our regular chats, or in Josh's case, our binge catch ups when I'm in Melbourne ;). Life is about fun and all you guys, Oxford and non-Oxford family, make the fun for me!

On a slightly different note, Professor Paul Glasziou, Dr Lea Sitkin and Professor Chris Del Mar all deserve recognition. They were all incredibly generous with their time during my application to Oxford. Without their advice and reassurance I wouldn't have made it. There are also other colleagues that contributed to my DPhil that I want to thank: Professor Richard Hobbs, Dr Jason Oke, Dr Ben Feakins, Ruth Davis, Dr Joe Lee, Dr Ben Goldacre, Dr Sarah Stevens, (soon-to-be Dr) Georgia Richards, and Dr Clare Bankhead.

A DPhil is a prolonged mental journey; avolition can insidiously creep in. Ridiculous as it may sound, music got me through at times, particularly after a heavy college lunch. I wish (need) to acknowledge Queens of the Stone Age, Arctic Monkeys (who released a new album during the last stretch of my writing!), David Bowie, LCD Soundsystem, The Strokes, The Cure, Alex Turner's solo work, Miles Kane, The Last Shadow Puppets, Fat Freddy's Drop, The War on Drugs, Sufjan Stevens, Blank Realm, Mac DeMarco, The Smiths, Franz Ferdinand, Cloud Control, The Libertines, Tame Impala, The Whitest Boy Alive, Royal Headache, Babyshambles, Pixies, and The White Stripes.

Finally, Katie Roche contributed enormously to this thesis; I am very grateful for all her support. It's cruel that she didn't have the opportunity to celebrate its completion.

To my future self, who I suspect will reread this too many times. Chill out.

Abstract

The aim of this thesis was to assess test use from primary care. Test use is an essential part of general practice, yet there is surprisingly little data exploring and quantifying its activity. My overarching hypothesis was that test use from primary care is sub-optimal, specifically that tests are overused (overtesting) - ordered when they will lead to no patient benefit, and underused (undertesting) - not ordered when they would lead to patient benefit. Previous metrics used to identify potential over and undertesting have been categorised into direct and indirect measures. Indirect measures take a population-level approach and are "unexpected variation" in healthcare resource use, such as geographical variation. Direct measures consider individual patient data and directly compare resource use with an appropriateness criterion (such as a guideline). In this thesis, I examined three indirect measures: temporal change in test use, between-practice variation in test use and variation between general practices in the proportion of test results that return an abnormal result. In chapter 3, I identified which tests have been subject to the greatest change in their use from 2000/1 to 2015/16 in UK primary care. In chapter 4, I identified the tests that had been subject to the greatest between-practice variation in their use in UK primary care. In chapter 5, I present a method to identify General Practices whose doctors order a lower proportion of tests that return a normal result. In chapter 6, I present a method to directly quantify over and undertesting; I conducted a systematic review of studies that measured the adherence of general practitioner's test use with guidelines. In chapter 7 I acknowledge that the use of guidelines to audit general practitioner's test use is flawed; guidelines are of varying quality and not designed to dictate clinical practice. In this chapter, I determine the quality and reporting of guidelines, the quality of the evidence underpinning their recommendations and explore the association between guideline quality and non-adherence. Overall, I have shown that most tests have increased substantially in use (MRI knee, vitamin D and MRI brain the most), there is marked between-practice variation in the use of many tests (drug monitoring, urine albumin and pelvic CT the most) and that some general practices order a significantly lower proportion of tests that return an abnormal result. I have also shown that there is marked variation in how often GPs follow guidelines, but guidelines based on highly quality evidence are adhered to significantly more

frequently. Lastly, in my Discussion chapter, I discuss the implications of my thesis, how it fits into the wider literature and an idea for a proposed step-wise approach to systematically identify overtesting.

Statement of Contributions

I certify that this thesis contains my own work. In instances where others have contributed to my word, their initials are used. The contributions of collaborators are summarised below:

- Sarah Stevens (SS) extracted the data from the Clinical Practice Research Datalink (CPRD), as described in Chapters 3 and 4.

- Ali Albasri (AA) conducted duplicate title and abstract screening and data extraction for the systematic review in Chapter 6. Ali Albasri also conducted duplicate AGREE II scoring for the guidelines in Chapter 7 and verified the GRADE assessments in Chapter 7.

- Brain Nicholson (BN) conducted duplicate title and abstract screening for Chapter 6.

- Nia Roberts (NR) reviewed the search strategy, which was used in Chapter 6.

Table of Contents

Chapter One Overview and Structure of Thesis 19

1.1 Overview 19

1.1.1 Aim of thesis 19

1.1.2 Overall Hypothesis 19

1.2 Thesis Structure 19

1.2.1 Chapter 2: Background 19

1.2.2 Chapter 3: The identification of tests with the greatest temporal increase in use 19

1.2.3 Chapter 4: The identification of tests with the greatest between-practice variation in use 20

1.2.4 Chapter 5: The identification of general practices that order a lower proportion of "abnormal" tests 20

1.2.5 Chapter 6: The assessment of test use against guidelines 21

1.2.6 Chapter 7: An assessment of the robustness of guidelines for auditing test use 21

1.2.7 Chapter 8: Discussion 22

1.3 Key Accomplishments 22

1.3.1 Publications 22

1.3.2 Grants 23

1.3.3 Teaching 23

1.4 Personal reasons why I did this thesis 23

Chapter Two Background 27

2.1 Chapter Aim 27

2.2 Chapter methods 27

2.3 Sustainability of healthcare systems 28

2.3 Increasing spending 29

2.5 Cost of tests 29

2.6 The role of tests in practice 30

2.7 Waste in healthcare 31

2.8 Drivers of waste in test use 32

2.9 Previous efforts to identify waste in healthcare 34

2.10 Primary care is the ideal setting to mitigate waste 38

2.11 Gaps in the literature and implications for future research 39

2.13 Chapter summary 39

Chapter Three: Identifying tests with the greatest temporal change in use 48

3.1 Publication statement 48

3.2 Introduction 48

3.3 Aim 48

3.4 Objectives 48

3.5 Methods 48

3.5.1 Study population 49

3.5.2 Included tests 49

3.5.3 Statistical analysis 50

3.6 Results 52

3.6.1 Identifying the test with the greatest increase in use 55

3.6.2 Identifying the test with the greatest decrease in use 58

3.6.2 Temporal patterns of use 59

3.7 Discussion 60

3.7.1 Summary of findings 60

3.7.2 Strengths and weaknesses in relation to other literature 60

3.7.3 Implications 62

3.7.4 How this chapter fits within wider thesis 65

3.8 Conclusion 65

3.9 Chapter Summary 65

Chapter Four: Identifying tests with the greatest between-practice variation 72

4.1 Publication statement 72

4.2 Introduction 72

4.3 Aim 72

4.4 Objectives 72

4.5 Methods 72

4.5.1 Study population 72

4.5.2 Included tests 73

4.5.3 Statistical analysis 74

4.6 Results 75

4.6.1 Identifying the specific test with the most variation in use 75

4.6.2 Variability of test use measured against rate of test use 84

4.7 Discussion 88

4.7.1 Summary of findings 88

4.7.2 Strengths and weaknesses in relation to other literature 88

4.7.3 Implications 91

4.7.4 How this chapter fits within my wider thesis 93

4.8 Conclusion 93

4.9 Chapter Summary 93

Chapter Five: Variation in diagnostic test requests and outcomes 99

5.1 Publication statement 99

5.2 Introduction 99

5.3 Aim 99

5.4 Objectives 99

5.5 Methods 100

5.5.1 Study design and setting 100

5.5.2 Data source 100

5.5.3 Statistical analysis 101

5.5.3 Data availability 102

5.6 Results 102

5.6.1 Variation in the crude proportion of abnormal CRP and ESR tests ordered 103

5.6.2 Random effects binomial regression model 104

5.7 Discussion 106

5.7.1 Summary of findings 106

5.7.2 Strengths and weaknesses in comparison with previous literature 107

5.7.3 Implications 108

5.7.4 How this chapter fits within wider thesis 110

5.8 Conclusion 110

Chapter Six: Auditing the use of diagnostic tests against guidelines: A systematic review 113

6.1 Publication statement 113

6.2 Introduction 113

6.3 Aim 114

6.4 Objectives 114

6.5 Methods 114

6.5.1 Protocol and registration 114

6.5.2 Search Strategy 114

6.5.3 Eligibility criteria 115
6.5.4 Study selection and data extraction 115
6.5.4 Statistical analysis 116
6.6 Results 117
6.6.1 Study selection and characteristics 117
6.6.2 Risk of bias 119
6.6.3 Percentage of diagnostic tests ordered in line with specific guideline recommendations 119
6.6.4 Underused tests 120
6.6.5 Overused tests 121
6.6.6 Variation of inappropriateness against the same guideline recommendation 123
6.7 Discussion 123
6.7.1 Summary of findings 123
6.7.3 Strengths and limitations in relation to other studies 124
6.7.2 Implications 126
6.7.6 How this chapter fits within my wider thesis 127
6.8 Conclusion 128
Chapter Seven An assessment of the robustness of guidelines for auditing test use 138
7.1 Publication statement 138
7.2 Introduction 138
7.3 Aim 139
7.4 Objectives 139
7.5 Methods 140
7.5.1 Study design 140
7.5.2 Guideline non-adherence 141
7.5.5 Assessment of evidence quality (GRADE) 143
7.5.6 Linear Regression models 147

7.5.6.3 Sensitivity analyses 148
7.5.6.4 Subgroup analysis 149
7.6 Results 151
7.6.1 Identifying and locating guidelines 151
7.6.2 Assessment of the quality and reporting of guidelines using AGREE II 152
7.6.3 Assessment of evidence quality (GRADE) 153
7.6.4 Association between guideline adherence and the quality and reporting of guidelines (AGREE II) 153
7.6.5 Association between guideline adherence and evidence quality (GRADE) 157
7.6.6 Sensitivity analysis 158
7.6.7 Subgroup analysis 165
7.7 Discussion 169
7.7.1 Summary of findings 169
4.7.2 Strengths and weaknesses in relation to other literature 170
7.7.3 Implications 173
7.7.4 How this chapter fits within the wider thesis 174
7.8 Conclusion 174
7.9 Chapter Summary 175
Chapter Eight: Discussion 180
8.1 Chapter Aim 180
8.2 Introduction 180
8.3 Summary of findings 180
8.4 Strengths and limitations in relation to previous research 182
8.5 Implications for policy and practice 187
8.5.1 A potential method to identify inappropriate diagnostic test use 187
8.5.1.1 Step 1 Identifying tests with the greatest temporal increase, between-practice variation and frequency of use 189

8.5.1.2 Step 2 Stakeholder involvement 192
8.5.1.3 Step 3 Identifying General Practices that order a lower proportion of "abnormal" tests 192
8.5.1.4 Step 4 Auditing test use against an appropriateness criterion 193
8.5.2 Limitations of the 4-step approach 194
8.6 Implications for future research 198
8.5.1 Validation of my approach 202
8.6 My reflections 204
8.6.1 My research reflections 204
8.6.2 My personal reflections 204
8.7 Conclusion 205
Appendix Chapter 1 211
1.1 DPhil publications 211
1.2 Non-DPhil Publications 211
1.3 NIHR Grant Award Letter 214
Appendix Chapter 3 215
3.1 ISAC protocol 215
3.2 Temporal patterns of use 266
3.3 RECORD Checklist 310
Appendix Chapter 4 322
4.1 Included tests 322
4.2 Adjusted and unadjusted coefficient of variation for each 44 tests (individual graphs) 324
4.3 RECORD Checklist 369
Appendix Chapter 5 381

5.1 RECORD checklist 381

5.2 The unadjusted and adjusted average proportion of abnormal CRP and ESR tests and total tests ordered per practice 393

Appendix Chapter 6 398

6.1 PRISM Checklist 398

6.2 Search strategy 400

6.3 Risk of bias 407

6.4: Table 1 410

6.5: Table 2 415

Appendix Chapter 7 423

7.1 AGREE II Scores 423

7.2 Specific guideline recommendations (that were too long for Appendix Table 7.1) 437

7.3: GRADE Profiles 439

7.4 References for Appendix table 7.2 and appendix table 7.3 443

7.5: 23 questions in AGREE II tool 445

7.6 AGREE II Formula 447

Appendix Chapter 8 449

8.1 Rank of 44 tests by temporal increase 449

8.2 Rank of 44 tests by frequency of use 451

8.3 Rank of 44 tests by between-practice variation 453

8.4 Combined rank (Temporal, between-practice variation and frequency of use) of all 44 tests 455

List of Figures

Figure 4.1 Rank order of variability of tests, adjusted for age, sex and deprivation. 77

Figure 4.2 Adjusted and unadjusted coefficient of variation. 82

Figure 4.3 Adjusted and unadjusted rates for C-reactive Protein (CRP) use 83

Figure 4.4 Variability and rate of tests.. 85

Figure 5.1 Variation in the proportion of abnormal CRP and ESR tests across Oxfordshire General Practices. 103

Figure 5.2 Funnel plot comparison of adjusted and unadjusted (raw) average proportion of abnormal CRP tests per GP practice. 104

Figure 5.3 Funnel plot comparison of adjusted and unadjusted (raw) average proportion of abnormal ESR tests per GP practice. 105

Figure 6.1 PRISMA flow diagram 118

Figure 6.2 Underuse of tests. 120

Figure 6.3 Overuse of tests. 122

Figure 7.1 Methods flow diagram 140

Figure 7.2 Inclusion and exclusion of measures of non-adherent testing 152

Figure 7.3 Scatter plot and median guideline quality and reporting (Overall AGREE II Score) and % guideline non-adherence 154

Figure 7.4 Distribution of scores for each Domain of the AGREEII tool 156

Figure 7.5 Scatter plot and median of each GRADE category and % guideline non-adherent test ordering 158

Figure 7.6 Low risk of bias studies: AGREE II scores and % guideline non-adherence 159

Figure 7.7 Moderate risk of bias: AGREE II scores and % guideline non-adherence 162

Figure 7.8 High risk of bias: AGREE II scores and % guideline non-adherence 164

Figure 7.9 Underuse: Scatter plot and median of guideline quality and reporting (Overall AGREE II Score) and % guideline non-adherence 166

Figure 7.10 Overuse: Scatter plot and median of guideline quality and reporting (Overall AGREE II Score) and % guideline non-adherence 168

List of Tables

Table 2.1 MEDLINE search strategy 27

Table 3.1 Age and sex adjusted rates of test use per 10,000 person years 53

Table 3.2 Average annual percentage increase in test use 56

Table 3.3 Tests with greatest increase, grouped by type (2000/1 to 2015/16) 58

Table 3.4 Tests with greatest increase, grouped by type (2004/5 to 2015/16) 58

Table 3.5 Tests that decreased in use from 2004/5 to 2015/16 58

Table 3.6 Tests that did not significantly change from 2004/5 to 2015/16 59

Table 4.1 Adjusted and unadjusted coefficients of variation 78

Table 4.2 Classification of all 44 tests by rate and CoV 86

Table 5.1 Practice characteristics 102

Table 5.2 Comparison of adjusted and unadjusted models 106

Table 7.1 Differences between GRADE for interventions and GRADE for Diagnostic accuracy[16,18–20] 145

Table 7.2 Results from Models A.1 to A.6: Domains of AGREEII and non- adherent testing 155

Table 7.3 Results from Model B 157

Table 7.4 Low risk of bias: Association between AGREE II Domains and guideline non-adherence 160

Table 7.5 Moderate risk of bias: Association between AGREE II Domains and guideline non-adherence 161

Table 7.6 Moderate risk of bias: Association between GRADE Score and guideline non-adherence 163

Table 7.7 High risk of bias: Association between AGREE II domains and guideline non-adherence 165

Table 7.8 Underuse: Association between AGREE II Domains and guideline non-adherence 166

Table 7.9 Underuse: Association between GRADE score and guideline adherence 167

Table 7.10 Overuse: Association between AGREE II Domains and guideline non-adherence 169

Table 7.11 Overuse: Association between GRADE score and guideline non-adherence 169

Table 8.1 Top 5 tests with greatest temporal change in use (2004/5 to 2015/16) 189

Table 8.2 Top 5 most frequently used tests (2015/16) 189

Table 8.3 Top 5 tests with most between-practice variation 190

Table 8.4 Top 10 tests for combined score 191

Chapter One Overview and Structure of Thesis

1.1 Overview

1.1.1 Aim of thesis

The aim of this thesis is to assess test use from primary care.

1.1.2 Overall Hypothesis

The overarching hypothesis of this thesis was that test use from primary care is sub-optimal.

1.2 Thesis Structure

1.2.1 Chapter 2: Background

In chapter 2, I describe the wider topic of healthcare sustainability. In this chapter I discuss waste in healthcare, the role and importance of tests in practice, the cost of test use, drivers of inappropriate test use, and previous efforts to a) assess healthcare resource use and b) quantify wasteful resource use.

1.2.2 Chapter 3: The identification of tests with the greatest temporal increase in use

In Chapter 3, my aim was to determine which tests had the greatest increase in use between 2000 and 2015. I chose 44 tests to study (28 laboratory, 11 imaging, and five miscellaneous tests); these tests were selected because they are commonly used tests, included in guidelines or in the Quality Outcomes Framework (QOF) or had to be grouped for our CPRD analysis. This list was refined with help from our patient and public involvement (PPI) group. I used data from CPRD to retrospectively analyse temporal trends in the use of tests from UK general practices. I calculated crude and age- and sex-standardised rates of test use. I then modelled temporal changes in standardised rates using joinpoint regression. I also conducted a sensitivity analysis from after the introduction of the QOF (2004–2015). I found that over the 16-year study period, MRI knee scans had the highest average

annual increase, followed by vitamin D tests and MRI brain scans. Vitamin D, MRI knee, and CT pelvis had the greatest increases after QOF. Vaginal swab was the only test in which the use fell significantly over the study period. I also found that renal function tests were the most commonly ordered test for most of the study period.

1.2.3 Chapter 4: The identification of tests with the greatest between-practice variation in use

In chapter 4 I identified tests with the greatest between-practice variation in use. I studied the same 44 tests used in chapter 3. I calculated the crude rates of use for each 44 tests from each general practice from April 1st 2015 to 31st March 2016. To adjust for differences in patient population between general practices, I constructed Poisson regression models. The Poisson regression model included three covariates: age, sex and deprivation. I constructed 44 Poisson models (one for each specific tests), which were used to produce an adjusted rate of test use for each general practice. The adjusted rates were used to calculate the adjusted coefficient of variation for each test. I ranked tests according to their adjusted coefficient of variation. The tests with the highest co-efficient of variation were non-illicit drug monitoring, followed by urine microalbumin, pelvic CT and then Pap smear. I also identified tests that had both a higher than average co-efficient of variation and a higher than average rate of use. Six tests met this classification: Clotting, Vitamin D, Urine Albumin, Prostate Specific Antigen (PSA), Bone profile, Urine MCS and C-reactive Protein (CRP)).

1.2.4 Chapter 5: The identification of general practices that order a lower proportion of "abnormal" tests

In chapter 5 I identified general practices that ordered a greater proportion of tests that returned a normal result (i.e. within the reference range), compared with their colleagues. I explored C-reactive Protein (CRP) and Erythrocyte Sedimentation Rate (ESR) tests ordered across Oxfordshire NHS General Practices. I calculated the proportion of tests that yielded an abnormal result for each practice and identified practices that had a proportion of abnormal CRP and ESR results 3 standard deviations

below the Oxfordshire mean. I estimated the adjusted average proportion of abnormal CRP and ESR tests that yielded abnormal results from each practice, after adjusting for differences in practice populations. These proportions were plotted against the total CRP and ESR requests per practice. I then constructed funnel plots to identify practices 3 standard deviations below the mean proportion of abnormal CRP and ESR tests. I analysed 143,745 CRP and 30,758 ESR requests from 69 practices. Twelve (17%) and seven (10%) practices were more than 3 standard deviations below the mean for CRP and ESR testing respectively. Two practices (3%) were below the 99.8% limit for both CRP and ESR ordering.

1.2.5 Chapter 6: The assessment of test use against guidelines

In chapter 6 I explored primary care physicians' non-adherence with diagnostic test guideline recommendations. I conducted a systematic review, which included primary studies that quantified either (or both) underuse, failure to order a test when indicated, or overuse, a test that was ordered when not indicated. 17 tests were underused >50% of the time. Of these, echocardiography (n=4 measures) was consistently underused (between 54% and 89%, n=4). There was large variation in the rate of inappropriate underuse of pulmonary function tests (38% to 78%, n=8). Eleven tests were inappropriately overused >50% of the time. Echocardiography was consistently overused (77%–92%), whereas inappropriate overuse of urinary cultures, upper endoscopy and colonoscopy varied widely, from 36% to 77% (n=3), 10% to 54% (n=10) and 8% to 52% (n=2), respectively.

1.2.6 Chapter 7: An assessment of the robustness of guidelines for auditing test use

In chapter 7 I aimed to explore the robustness of guidelines as a tool to directly audit the appropriateness of using tests. I assessed guidelines used to determine the appropriateness of ordering tests in the systematic review reported in chapter 6. I also assessed the quality of evidence supporting these guideline recommendations. We (Ali Albasri also independently did this) used the Appraisal of Guidelines for Research & Evaluation (AGREE) II tool to assess the proficiency of guideline development and Grading of Recommendations Assessment, Development, and Evaluation (GRADE)

to examine the quality of evidence underpinning these recommendations. Finally, I constructed linear regression models to determine the association between guideline non-adherence (data from Chapter Six) and a) the AGREE II score and b) the GRADE score. The AGREE II scores for the guidelines were varied but largely suboptimal (median 4.5 (IQR: 3.5-5)). Most (n=45, 73%) of the evidence underpinning guideline recommendations was graded as very low, with seven graded as low (11%), and ten graded as high (16%) (none was graded as moderate). There was no significant association between the AGREE II score and non-adherent testing (P = 0.09). There was, however, a significant association between the GRADE score and non-adherence. Guideline recommendations based on low-quality and very low-quality evidence had 38% (95%CI 10% to 65%, P = <0.01) and 24% (95%CI: 4.5% to 43%, P = 0.02) more non-adherent testing, compared with recommendations based on high quality evidence. To directly audit the appropriateness of test ordering, guideline development needs to be robust and transparent and recommendations need to be supported by high-quality evidence. Currently, few guidelines meet this requirement.

1.2.7 Chapter 8: Discussion

In this chapter, I summarise my findings, how they add to the existing literature and the limitations of my research. I also discuss the implications of my research for practice, specifically I explore an approach to identify unnecessary test use in primary care, its limitations and the required future research.

1.3 Key Accomplishments

1.3.1 Publications

At the time of submission, three papers from my DPhil research have been published in peer-review journals.[1,2,3] A fourth paper has been accepted by The BMJ and the has been provisionally accepted at BMC Medicine.

Additionally, I have had the opportunity to work on research beyond my DPhil. This included papers which I led, as well as collaboration with other researchers. Since arriving in the department I have

accumulated 17 further publications. Seven of these are first author publications, including one in The BMJ (*BMJ* 2018;361:k2387). Chapter 1 Appendix lists my DPhil and non-DPhil publications.

1.3.2 Grants

My DPhil was funded by The Clarendon Fund. This is a fantastic scholarship, but unfortunately comes without dedicated research costs. To support chapters 3 and 4, I was successful in obtaining funding from the National Institute of Health Research (NIHR) School of Primary Care (External funding round: £92,000, Award number 386) and the Primary Care Research Trust (£10,000). Chapter 1 Appendix lists the first page of the NIHR award.

1.3.3 Teaching

Through the Centre for Evidence-Based Medicine (CEBM), I have had the opportunity to teach on the Masters of Evidence-Based Healthcare. I lectured on the 2017 and 2018 Evidence-Based Diagnosis and Screening Module. I also was a course tutor on the following modules: The Practice of Evidence-Based Health Care, Introduction to Study Design and Research Methods and Evidence-Based Diagnosis and Screening. I am grateful to CEBM and the Nuffield Department of Primary Care Health Sciences for allowing me to attend the Teaching Evidence-Based Medicine module, which prepared me to teach.

1.4 Personal reasons why I did this thesis

My experiences as a junior doctor motivated me to undertake this thesis. As a junior doctor, I worked under many Consultants in all aspects of medicine and was struck by: a) the volume of tests ordered, b) the little added value of many tests, and c) the unexpected consequences and harms of tests. This experience, coupled with the knowledge that healthcare systems around the world were struggling to maintain financial sustainability, led me to try to identify wasteful use of tests. Several clinical examples particularly influenced me, I have listed them under the broad headings of medicolegal fear, incidental findings, and overdiagnosis.

- Medicolegal fear
    - I was working night shifts and was part of the immediate response team that were called to see a 12-year old girl with chest pain and tachycardia. She was in hospital for a psychiatric condition, and the clinical impression from the senior doctor running the response was that her symptoms were unlikely to represent a sinister underlying condition. However, he concluded that Computerised Tomographic Pulmonary Angiography (CTPA) was required because: "if someone looks at the notes and sees a patient in hospital with chest pain and tachycardia they will ask why we didn't do a CTPA". The test was ordered, although the doctor was convinced clinically that the patient did not require one, and that the result would be normal. The CTPA report confirmed that no pulmonary embolism (PE) was present.
- Incidental findings:
    - I was involved in the care of a patient with an incidental brain lesion found on a CT scan. She was asymptomatic and young (~30); there was uncertainty around the diagnosis, which lead to a neurosurgical biopsy. Intraoperatively she bled, leaving her with a permanent physical disability and unlikely to work again. The biopsy showed a benign growth. This was obviously a terrible outcome and I couldn't help but ruminate about the initial scan; it had not been indicated and if it hadn't been done perhaps none of the above would have happened.
- Overdiagnosis:
    - I was working in the Emergency Department when an 18-year-old girl presented with a headache. She had presented two days earlier after fainting and knocking her head. She had chronically low blood pressure, was dehydrated at the time, and had fainted many times in the past. She had been reassured and sent home. I saw the patient when she presented again. She was well oriented, without any amnesia or change in behaviour, and the physical examination was completely normal. After discussing the case with my consultant, he suggested a CT brain scan – mainly because the patient

had represented. Surprisingly, she had a small intraparenchymal haematoma. We referred her to the neurosurgical department who admitted her for two days and then discharged her, without any further intervention. When I discuss this case with most clinicians, they speak of a relief in finally uncovering the diagnosis. I understand this, and it is a relevant point, but I don't think it is as simple as that. The patient was entirely asymptomatic aside from a headache, the diagnosis did not lead to any further intervention, and she recovered. She did not seem particularly reassured after she had been given the diagnosis. I am not convinced her outcome would have been any different without the definitive diagnosis. Obviously most patients and clinicians would like to know if they have an intracerebral haematoma and it was advantageous that we diagnosed this. However, this patient was the first that led me to question the utility and, in some cases, futility of a diagnosis. I began to reflect on what makes a diagnosis useful and how increasingly sensitive tests are likely to beget an increase in less useful diagnoses. I concluded that tests should not be ordered unless we suspect that we can reach a diagnosis for which an intervention would be effective. It also led me to believe that there is a need to understand more broadly the natural progression of disease.

All of these clinical examples question the appropriateness of tests and also outline some of the tangible harms I have seen patients being subjected to because of the use of tests.

References

1 O'Sullivan JW, Heneghan C, Perera R, *et al.* Variation in diagnostic test requests and outcomes: a preliminary metric for OpenPathology.net. *Sci Rep* 2018; **8**: 4752.

2 O'Sullivan JW, Albasri A, Nicholson B, *et al.* Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open* 2018; **8**: e018557.

3. O'Sullivan JW, Albasri A, Koshiaris C, et al. Diagnostic test guidelines based on high-quality evidence had greater rates of adherence: a meta-epidemiological study. J Clin Epidemiol 2018;103:40–50. doi:10.1016/j.jclinepi.2018.06.013

Chapter Two Background

2.1 Chapter Aim

The aim of this chapter is to describe the rationale for this thesis. Specifically, I discuss the unsustainable trajectory most healthcare systems currently face, including the NHS. I explore the amount of wasteful spending in healthcare budgets and previous efforts to identify profligate resource use.

2.2 Chapter methods

This chapter is supported by a multifaceted search strategy. The three key steps are:

1. A structured, formal search of the MEDLINE database via the PubMed interface. Table 2.1 outlines the MEDLINE search strategies used.
2. Opportunistic sampling: Over my three years of research within this field, I have collected the pertinent literature. Some of which underpins this chapter.
3. Forward and backward citation searches were also performed on the relevant literature.

Table 2.1 MEDLINE search strategy

Topic	**Search strategy**
Sustainability of healthcare systems	(spend* OR expenditure) AND healthcare AND sustainability[Title/Abstract]
Wasteful diagnostic test use	"Unnecessary Procedures"[Mesh] AND "Diagnostic Tests, Routine"[Mesh]
	"Guideline Adherence"[Mesh] AND "Diagnostic Tests, Routine"[Mesh]

2.3 Sustainability of healthcare systems

Healthcare systems around the world are struggling to remain financially sustainable.[1] The US Medicare deficit is estimated to reach $660 billion by 2023,[2] while in the UK, services are being rationed to try and meet the required £22 billion in efficiency savings.[3] There are similar estimates in both other high-income countries, and low and middle-income countries.[4]

Concerns surrounding the sustainability of the NHS have emerged due to a mismatch between NHS funding and expenditure.[5] Increasing spending, coupled with historically low funding (as a proportion of GDP), has led to concerns about an 'unbridgeable' funding gap.[6] Financial pressures has led to rationing of services[7] and longer waits to see GPs,[8] at Accident and Emergency (A&E) departments,[9] and for elective operations.[10] It has also led to closures of GP surgeries[11] and A&E departments,[12] and increased workload pressures.[13] Concerns regarding NHS sustainability have led to a House of Lords inquiry, which concluded that: "leaders in the NHS seem to be incapable of driving the much needed change in levels of productivity" and "unwarranted levels of variations in patient outcomes are unacceptably undermining the effectiveness and efficiency of the NHS".[14] It also reinforced the need for a "tax-funded, free-at-the-point-of-use NHS".[14]

In response to calls for efficiency savings, the NHS have made commendable, but ultimately insufficient improvements.[15] The Nuffield Trust estimates that annual savings of 4% are required to reach £22 billion in efficiency savings.[15] From 2011 to 2015, the NHS managed to reduce its expenditure by around 2%.[15]

The NHS, like many other healthcare systems, is in a precarious situation. In the absence of increasing funding, large increases in efficiency are required to match NHS funding and expenditure, and thus ensure sustainability.

2.3 Increasing spending

The NHS budget is predicted to grow by 1.2% in real terms annually, while spending is expected to increase by 4% a year (above inflation).[5] NHS spending has seen a rapid rise in the past 15 years. From 2003 to 2016, net expenditure has increased by over £50bn – an 80% rise.[16] Spending has also increased in most other healthcare systems around the world; in the USA spending increased by 5.8% annually from 2014, reaching $3.2 trillion in 2015.[17] Increased spending leaves healthcare systems in a perilous financial situation, but also has wider consequences. Countries have finite resources - what Nobel Prize winning economist Eleanor Ostrom labelled "common pool resources"[18] - thus greater spending on health drains resources from other public services. More money spent on health leaves less money for schools, infrastructure and public transport, jeopardising the capacity of other public services.

Over the last 40 years, healthcare spending as a percentage of gross domestic product (GDP) has increased in all high-income countries.[19] In the US, healthcare spending as a percentage of GDP increased from 6.2% in 1970 to 17.2% in 2016,[20] a 176% increase. Similar increases were seen in the UK; healthcare spending as a percentage of GDP increased 145% over the same time period; from 4.0% to 9.7%.[20] Spending as a percentage of GDP is expected to double in the UK by 2065.[21] Notably, greater spending doesn't necessarily lead to improved patient outcomes; the US spends the most on healthcare (as a proportion of GDP), but has the worse outcomes, compared to other high-income countries.[19] In recent years, life expectancy in the US is actually decreasing.[19]

2.5 Cost of tests

Lord Carter of Coles' 2006 report on NHS pathology estimated that tests influence around 70% of clinical decisions.[22] Despite their importance, there is a paucity of evidence quantifying test use expenditure. Previous estimates are outdated and only included laboratory tests. For instance, the 2006 Lord Carter report concluded that between £2 and 3 billion of the 2006 NHS budget (~4% of the £76.4 billion 2006 NHS budget)[23] was consumed by pathology tests alone.[22] Given that test use has increased substantially since then, it is likely that test expenditure has also increased.

Furthermore, these estimates only include direct costs of test use. Direct costs refer to the total cost to deliver the test. For a blood test for instance, direct costs refer to the cost of the phlebotomist and their equipment, transport of the sample to the laboratory, the scientist's time to run the sample and the running costs of the laboratory machine.[24] Direct costs do not account for the time taken by the requesting clinician to review and action the test result. Furthermore, tests can often have knock-on resource consequences. For instance, incidental imaging findings can lead to further consultations, imaging and in some cases, treatment (sometimes this benefits the patient, other times it causes more harm than good).[25] Similarly, excessive biological monitoring of chronic disease can lead to spuriously positive results.[26] For instance, research has shown that the greater the number of cholesterol tests, the greater number of false positive results.[26] False positive results can lead to further consultations, tests and treatment changes. This has sometimes been called Ulysses syndrome. It should be noted that Ulysses syndrome has traditionally referred to an atypical set of depressive symptoms thought to be related to stress of modern migration.[65] However, it has also been used to describe patients who "find themselves caught in a web of further investigations, referrals, and sometimes treatment before finally being recognised as healthy, which they were in the first place."[66] It is this definition which I refer to above: the stress that comes with a false-positive diagnosis. Although the definition related to the stress of migration is perhaps better known, Ulysses syndrome related to a false-positive diagnosis was first described almost 50 years ago; it was described in 1972 as: "Mental and physical disorders which follow the discovery of a false positive result".[67]There are also non-financial costs of tests to consider. Some imaging tests subject patients to huge amounts of radiation,[27] Computer Tomography (CT) contrast also risks nephropathy[28] and endoscopic tests carry life-threatening risks.[29] Furthermore, the discomfort many patients go through with venepuncture is not trivial for many.

2.6 The role of tests in practice

Despite their costs and potential harms, tests are hugely important to medical practice. The different roles of tests can be broadly categorised into monitoring, screening and diagnostic.[30] Monitoring tests

are defined as “repeated testing aimed at guiding and adjusting the management of a chronic or recurrent condition”.[31] Monitoring tests can be imaging or endoscopic, but are commonly biochemical or haematological tests. A common example of a monitoring test is glycated haemoglobin (HbA1c). HbA1c is used to measure the control of diabetes. However, note that HbA1c can also be used as a diagnostic test (initial test). The role of monitoring tests is to detect a deterioration or improvement in a patient with an established disease.[31] For a monitoring test to be useful, a change in a test result should predict a change in clinical status. For example, an elevated B-type natriuretic peptide (BNP) in a patient with heart failure predicts worsening morbidity and earlier mortality.[32]

Tests can also be used to screen for disease. Screening tests are performed on asymptomatic patients or symptomatic patients undergoing testing for an unrelated reason.[33,34] They aim to detect disease before it emerges clinically, thus offering an opportunity to intervene early. Screening tests can be broadly classified into ‘organised’ and ‘opportunistic’. Organised screening programs are offered to eligible populations in a structured and regular way. For example, breast cancer screening with mammography. In the UK, the NHS offers 11 organised screening programs (abdominal aortic aneurysm (AAA), bowel cancer, breast cancer, cervical cancer, diabetic eye disease, fetal anomalies, infectious diseases in pregnancy, physical examination of neonates and infants, neonatal blood spot for hereditary metabolic diseases such as phenylketonuria, neonatal hearing, and sickle cell and thalassaemia).[35] Opportunistic screening occurs when a test is offered to a patient who presents to a health professional for an unrelated reason. For instance, if a 65-year-old man presents to the GP with flu-like symptoms and additionally mentions his concerns about prostate cancer (without any signs or symptoms of prostatic disease), ordering a Prostate-Specific Antigen (PSA) test would represent opportunistic screening.

Lastly, diagnostic tests are those ordered for symptomatic patients, to assist a clinician in ruling in or out a disease and therefore aiding in reaching a diagnosis.[36]

2.7 Waste in healthcare

To improve the sustainability of healthcare systems, much effort has focused on identifying waste. Waste in healthcare can occur through the delivery of clinical care and/or administrative processes.[37] Waste in clinical care can be thought of as overuse, "the provision of medical services that are likely to cause more harm than good",[38] or underuse, the failure to deliver effective tests and treatments.[39] Administrative waste represents managerial decisions that lead to decreased efficiency of health service delivery or expenditure that does not improve services.[37] Clinical care is estimated to be the most significant contributor to waste in healthcare delivery[37] and thus will be the focus of this thesis.

Overuse wastes finite healthcare resources, but also risks avoidable harms to patients. For instance, the overuse of diagnostic tests put patients at risk of harms such as radiation exposure, pain, and adverse outcomes (e.g. contrast nephropathy).[28] Overuse is estimated to represent around 25-30% of medical care,[40] and is thought to be increasing.[38]

Underuse, depriving patients of the appropriate tests and treatments, can lead to delays in diagnosis and effective treatment, often resulting in adverse patient outcomes and further healthcare costs.[41,42] Underuse is common in low and middle-income countries through the lack of healthcare access and availability of effective interventions.[39] For instance, around the world it is estimated that over 400 million people lack access to basic healthcare.[43] However, underuse is also common in high-income countries, largely through the lack of delivery of effective care.[39,44]

Waste is the opposite of "right care"[1], also known as "appropriate care".[30] Right care is defined as "care that weighs up the benefits and harms, is patient-centred, and is informed by evidence, including cost effectiveness".[1] All healthcare systems strive to achieve right care, but estimates indicate that waste is prevalent through most healthcare systems. A 2017 report from The Organisation for Economic Co-operation and Development (OECD) estimated that around a fifth of healthcare spending is wasteful.[37] The challenge remains to mitigate waste without jeopardising appropriate care.

2.8 Drivers of waste in test use

Numerous drivers of waste in test use have been described. There are two reports which have systematically grouped these drivers: The Organisation for Economic Co-operation and Development (OECD) report on "Tackling wasteful spending on health",[37] and The Lancet Right Care series.[45]

The 2017 OECD report divides drivers of healthcare waste into four categories:

1. 'Healthcare professionals do not know better': A healthcare professional's lack of knowledge and cognitive biases contribute to them delivering inadequate care, not in line with best practice. The dissemination of new research and knowledge remains a challenge. Primary care centres are particularly susceptible to this deprivation of new knowledge as they often exist in isolation without structures in place to keep up to date.
2. 'Healthcare professionals cannot do better': Structural flaws exist within the healthcare system that prevent healthcare professionals providing optimal care. This may be due to insufficient funding or poor management and organisation.
3. 'Healthcare professionals could stand to lose by doing the right thing': This occurs when healthcare professionals are incentivised to provide care that may not be in line with the goals of the healthcare system. For instance, in many healthcare systems, clinicians economically benefit through the provision of care, regardless of the added value of the care. It also includes medicolegal fears of "missing something".
4. 'Healthcare professionals intentionally generate waste, for self-interest'. This driver is related to the third but refers explicitly to fraud and corruption.

The Lancet Right Care series also describe drivers of 'poor medical care'.[46] The Lancet classification, as opposed to the OECD's, defines drivers of both over and underuse of health care resources. They define four categories:

1. Knowledge, beliefs, assumptions, bias and uncertainty: This driver maps closely to the OECD driver 'Health professionals do not know better'. The Lancet, however, discusses more explicitly patient's and clinician's misconceived beliefs that contribute to the delivery of poor care. Patient beliefs they describe include: Testing is accurate (poor understanding of error

rates and other limitations), more care is better and anxiety surrounding uncertainty.[46] There are a range of clinician biases, beliefs and habits that drive poor care including: when the evidence contradicts their training or experience, a 'better-to-know' bias (a clinician's desire to know definitively via a test), improper weighing of relative versus absolute risk and technological enthusiasm.[46] Medicolegal fear is also included under this category.

2. Money, finance and organisation: This driver maps closely to the OECD driver 'Healthcare professionals cannot do better'. The Lancet description also includes poor health coverage as a driver of underuse. Furthermore, this driver is related to the OECD driver 'Healthcare professionals intentionally generate waste, for self-interest', where explicit fraud is acknowledged as a driver of waste in healthcare.
3. Economic Incentives: This driver maps closely to the OECD driver 'Health professionals could stand to lose by doing the right thing'. The Lancet also describes patients' behaviour as a response to economic advantage. For instance, they describe varying levels of insurance enabling access to different levels of care. They note that greater access and more tests does not always lead to better outcomes and is a likely driver of waste.
4. Power and human relationships. This driver is not described in the OECD classification. The Lancet authors describe the quality of the patient-clinician relationship contributing to both over and underuse. They state that 'the absence of mutual respect and trust' can lead to inadequate exchange between patients and clinicians (such as a patient history), which can facilitate suboptimal care.[46]

The OECD and Lancet classifications comprehensively map the drivers of waste in healthcare. They are closely aligned; however the Lancet classification is preferable because of its focus on both over and underuse and its classification of 'power and human relationships' as a driver of waste.

2.9 Previous efforts to identify waste in healthcare

Previous metrics used to identify potential waste in healthcare resource use have been broadly classified into direct or indirect measures. Indirect measures take a population-level approach and

have been defined as "unexpected variation" in healthcare resource use; variation that persists despite adjustment for differences in patients or population characteristics between regions or over time.[47] The most widely studied indirect measure is geographical variation. Geographical variation in healthcare resource use - variation not attributable to differences in patient characteristics - has been documented within high,[48,49] middle[50] and low-income countries.[51] It has also been documented between countries and healthcare systems.[52] Temporal change in resource use is another indirect measure. Measures of geographical and temporal variation do not directly assess the indications for care and they therefore represent, at best, a potential marker of over and underuse. It is not possible to draw conclusions about the appropriateness of resource use when interpreting temporal or geographical variation in isolation. Evidence of variation is only used to identify outliers, which may benefit from further investigation. For example, a study quantifying geographical variation may identify a general practice that orders chest x-rays at a rate greater than the upper confidence interval of the national average. This general practice may or may not be overusing chest x-rays – this cannot be deduced from the study. However, the existence of geographical variation has identified a general practice that would benefit from a direct assessment of the appropriateness of their test ordering, such as an audit. A direct assessment can then quantify if the outlying practice is overusing chest x-rays.

Direct assessment of over and underuse of healthcare resources requires an examination of individual patient data (IPD), which can be used to determine if the ordering of the health care resource (for instance, a test) was appropriate. To do this, IPD should be compared with a high-quality 'appropriateness criteria'; guidelines have been the most commonly used 'appropriateness criteria.'[30,53–55] Some previous studies have used the term "low-value care"[64] to describe overuse. While the term "low-value care" may be considered by many to be adequate, I find this term ambiguous. I therefore preferred to be more explicit and, in this thesis, I use the phrase: a direct assessment of the appropriateness of test use against guidelines. Regardless of the nomenclature, the use of guidelines to directly audit the appropriateness of test use is flawed, given the varying quality of guidelines. Furthermore, in the absence of explicit guidelines, some organisations develop their own appropriateness criterion. A common method to achieve this is the RAND-UCLA

appropriateness method. This method reaches a recommendation through a mixture of scientific evidence and expert opinion.[56] Notably, many guidelines incorporate the RAND-UCLA appropriateness method (or something similar, such as the Delphi method) when generating their recommendations.

In the UK, both indirect and direct measures of waste have been documented. The NHS atlases of variation describe the marked variation in the use of tests and treatments throughout the UK. The 2nd Atlas of Variation in NHS Diagnostic Services in England [49] specifically describes the variation in test use between clinical commissioning groups (CCGs). The atlas describes the variation of around 30 tests between the 209 CCGs in England. Because the unit of analysis is CCGs, the atlas identifies CCGs that have a rate of test use above or below the 99.8% confidence interval of the national average. The atlas shows that marked variation in diagnostic services exists within England.

More broadly, Lord Carter conducted an independent report for the UK Department of Health on the operational productivity and performance of English NHS acute hospitals.[57] This report focuses on many aspects of acute hospital activity – both clinical and non-clinical expenditure, including pathology and imaging services. Throughout the report they acknowledge the difficulty in directly assessing the appropriateness of resource use in NHS hospitals and, instead, focus exclusively on documenting geographical variation in various activity measures between acute NHS hospitals. To compare the expenditure of an activity across hospitals they measure the proportion of total operating expenditure the activity consumes for each hospital. They then compared this proportion across hospitals and report the dispersion and highlight hospitals above and below the 95th and 5th centile (from the median). For instance, they reported that the median NHS English hospital spends 1.57% of its total operating expenditure on pathology costs. They also reported a 176% variation between the 5th and 95th centile and highlight one hospital beyond the 95th centile and one below the 5th centile. Furthermore, the same report highlights the 52% variation in the cost of imaging as a proportion of operating expenditure (median 3.6%) and highlighted hospitals beyond the 95th and 5th centile respectively.

One previous academic study quantified temporal and geographical variation in UK test use.[58] Unlike the NHS atlases, this study used tests, rather than CCGs, as the unit of analysis and thus identified tests, rather than regions that are subject to much variation (temporal and geographical). Busby et al studied the use of 29 laboratory tests from 2005 to 2009 in a random sample of 100,000 patients.[58] This study used data from the UK's CPRD and reports the laboratory tests with the greatest increase in use and also the tests subject to the most geographical variation.

Studies directly measuring waste in healthcare resource use are increasingly common. A 2013 review quantified the over and under use of secondary care laboratory tests; they synthesised data from 42 primary studies, all of which used guidelines to determine the appropriateness of test use. These primary studies assessed the IPD of patients that had undergone laboratory tests and compared them with the respective guidelines. Furthermore, the Lancet *Right Care* series includes detailed reviews on both over and underuse of medical services around the world.[38,39] They report overuse in high, middle and low-income countries and include studies that describe direct measures of overuse across all specialities and in the use of both tests and treatments. Conversely, they also note the remarkable lack of research directly measuring underuse of healthcare resources. From the available data, they conclude that underuse is present 'across countries, regardless of the payment model', in both rural and tertiary settings and in high, middle and low-income countries.[39]

Most reports that quantify waste in healthcare systems use a mix of indirect and direct measures. The Institute of Medicine National Roundtable (USA) on healthcare quality[59] stated that underuse and overuse, "occur in small and large communities alike, in all parts of the country, and with approximately equal frequency in managed care and fee-for-service systems of care." To reach this conclusion, they acknowledged indirect markers of waste, most notably geographical variation. In regards to geographical variation they reported that "these (geographical variation) and other relevant issues may be causal or explanatory factors leading to a better understanding of quality problems; that is, they will be related to specific underuse, overuse, or misuse problems".[59] They concluded, using indirect and direct evidence, that "The evidence is compelling. Millions of Americans are not reached

by proven effective interventions that can save lives and prevent disability. Perhaps an equal number suffer needlessly because they are exposed to the harms of unnecessary health services".

Lastly, a 2012 study presented a novel approach to identify health care services of "low-value".[60] This review set out to identify services (tests and treatments) that were currently recommended (approved for reimbursement on the Australian Medicare Benefits Schedule), but their use was not supported by evidence. They identified tests and treatments where the underlying evidence did not support Australian reimbursement practice.

2.10 Primary care is the ideal setting to mitigate waste

Primary care represents the ideal setting to mitigate wasteful spending. It accounts for 90% of UK consultations[13] and around 55% of care provided in the USA.[61] This volume of care means that small improvements in care can translate to substantial benefits. Improvements would represent an increase in the use of tests when the benefits outweigh the harms. For instance, the appropriate use of echocardiography to confirm or refute a diagnosis of heart failure.[30] Improvements in test use would also represent a reduction in the use of tests when they cause more harm than benefit. These improvements would lead to benefits for patients and, more broadly, the healthcare system. Benefits to patients include a reduction in diagnostic delays and thus earlier initiation of the appropriate treatment. Improvements would also lead to a reduction in the number of patients subject to the risks of tests, for no benefit. For instance, a reduction in CT scans for non-red flag low back pain would reduce the number of patients subjected to unnecessary radiation. Patients would also benefit, indirectly, through the liberation of previously wasted resources.

Furthermore, in the UK the number and length of primary care consultations have increased significantly.[13] Between 2007 and 2014, the crude annual consultation rate per person increased by 11% and the mean duration of GP consultations (face-to-face) increased by 7%, from 8.7 minutes to 9.2 minutes. It has been suggested that the noted increase in primary care test use may be a direct response to these workload changes, with tests reportedly being used for 'strategic, non-medical reasons', such as reassuring patients and ending consultations.[62] Improvements in test use, both an

increase in appropriate test use (reducing underuse) and a decrease in inappropriate test use (reducing overuse), may reverse the unsustainable trajectory of primary care consultations. The appropriate use of tests could lead to a reduction in consultations through the early confirmation of a diagnosis and by reducing follow-up appointments to discuss tests that did not aid management.

Lastly, the complexity and breadth of patients presenting to primary care makes appropriate use of tests a challenge. The combination of undifferentiated symptoms, a low prevalence of serious disease, a high degree of symptom overlap between serious and benign conditions, patients with multiple complaints, and psychological or social distress manifesting with somatically all complicate reaching a diagnosis and can lead to sub-optimal test use.[63] This difficulty is likely to worsen, as people live longer and the number of people living with co-morbidities increases.

2.11 Gaps in the literature and implications for future research

Despite their frequency of use, there is a paucity of literature that assesses test use from primary care. There is one study that quantifies temporal change and geographical variation in UK primary care test use.[58] However, this study is outdated, examines a small group of select tests (laboratory tests only), was conducted in a small sample size and uses sub-optimal methods. There is no research that highlights UK general practices that may order tests too often or not enough. Furthermore, there is only one study, from a US group, that directly quantifies the appropriateness of test use against guidelines. This study did not focus on primary care test use. Lastly, no research has examined the robustness of guidelines as a tool to audit test use against, nor is there any literature that addresses the relationship between diagnostic test guideline quality and guideline concordance.

Furthermore, there is currently no national or regional policy strategy to identify unnecessary test use from primary care. Nor is there any research to address this. The research in my thesis will address these gaps in the evidence.

2.13 Chapter summary

- Healthcare systems around the world are struggling to remain financially sustainable. The NHS is rationing services to meet calls for £22 billion in efficiency savings. Questions surrounding the sustainability of the NHS have emerged among a mismatch between NHS funding and expenditure. Globally, healthcare spending continues to increase.
- The identification and mitigation of healthcare waste is currently a priority of many healthcare systems, including the NHS.
- Primary care, given the volume of care it provides, is the ideal setting to mitigate waste.
- Despite their importance (70% of clinical decisions are influenced by tests) and expenditure (test use was estimated to cost the NHS £2-3 billion in 2006) there is remarkable little evidence exploring primary care test use.

References

1 Kleinert S, Horton R. From universal health coverage to right care for health. *Lancet* 2017; **390**: 101–2.

2 Fisher ES, Bynum JP, Skinner JS. Slowing the growth of health care costs--lessons from regional variation. *N Engl J Med* 2009; **360**: 849–52.

3 Alderwick H, Robertson R, Appleby J, Dunn P, Maguire D. Better value in the NHS The role of changes in clinical practice. 2015.

4 Scrutton J, Holley-Moore G, Bamford S-M. Creating a Sustainable 21st Century Healthcare System. 2015 http://www.ilcuk.org.uk/images/uploads/publication-pdfs/Creating_a_Sustainable_21st_Century_Healthcare_System.pdf.

5 The King's Fund. The NHS budget and how it has changed. London, 2017 https://www.kingsfund.org.uk/projects/nhs-in-a-nutshell/nhs-budget.

6 Hopson C. The gap between funds and delivery is a chasm in the NHS: something has to give. Guard. 2016.

7 Iacobucci G. Exceptional requests for care surge as rationing deepens. *Bmj* 2017; **358**: 3188.

8 Appleby J. What happens when targets for waiting times change? *BMJ* 2017; **357**: 1–5.

9 Hawkes N. NHS data show A&E patients waiting longer and big rise in GP flu consultations. *Bmj* 2018; **188**: k188.

10 Iacobucci G. NHS cancels planned surgery and outpatient appointments in response to winter crisis. *Bmj* 2018; **19**: k19.

11 Iacobucci G. GPs call for ballot on mass closure of practice lists. *Bmj* 2017; **2494**: j2494.

12 Campbell D. A&E units, GP surgeries and walk-in centres to close as cash crisis bites. Guard. 2017. https://www.theguardian.com/society/2017/nov/15/ae-units-gp-surgeries-and-walk-in-

centres-to-close-as-cash-crisis-bites.

13 Hobbs FDR, Bankhead C, Mukhtar T, *et al.* Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14. *Lancet* 2016; **387**: 2323–30.

14 Select Committee on the Long-term Sustainability of the NHS. The Long-term Sustainability of the NHS and Adult Social Care. 2017 https://www.publications.parliament.uk/pa/ld201617/ldselect/ldnhssus/151/151.pdf.

15 Gainsbury S. Nuffield Trust: Feeling the crunch: NHS finances to 2020. 2016 https://www.nuffieldtrust.org.uk/research/feeling-the-crunch-nhs-finances-to-2020.

16 NHS Confederation. Key statistics on the NHS - NHS Confederation. Resources. 2016. http://www.nhsconfed.org/resources/key-statistics-on-the-nhs (accessed Aug 19, 2016).

17 Martin AB, Hartman M, Washington B, Catlin A. National Health Spending: Faster Growth In 2015 As Coverage Expands And Utilization Increases. *Health Aff* 2017; **36**: 166–76.

18 Ostrom E. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge: Cambridge University Press, 1990 DOI:DOI: 10.1017/CBO9780511807763.

19 Schneider EC. Mirror, Mirror 2017: International Comparison Reflects Flaws and Opportunities for Better US Health Care. *Med Benefits* 2017; **34**: 8.

20 Organisation for Economic Co-operation and Development (OECD). Health spending (indicator). 2018. DOI:10.1787/8643de7e-en.

21 Appleby J. Is the NHS financially sustainable ? 2016 https://www.nuffieldtrust.org.uk/resource/is-the-nhs-financially-sustainable.

22 Lord Carter of Coles. Report of the review of NHS pathology services in England. 2006. http://webarchive.nationalarchives.gov.uk/20130107105354/http:/www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_091984.pdf (accessed Jan

5, 2018).

23 The King's Fund. How much has been spent on the NHS since 2005? 2010. https://www.kingsfund.org.uk/projects/general-election-2010/money-spent-nhs.

24 National Institute for Health and Care Excellence. Routine preoperative tests for elective surgery. 2015. https://www.nice.org.uk/guidance/NG45/documents/guideline-appendices-13 (accessed Jan 15, 2018).

25 Lumbreras B, Donat L, Hernández-Aguado I. Incidental findings in imaging diagnostic tests: a systematic review. *Br J Radiol* 2010; **83**: 276–89.

26 Glasziou PP, Irwig L, Heritier S, Simes RJ, Tonkin A. Cholesterol Levels : Measurement Error or True Change ? *Ann Intern Med* 2008; **148**: 656–61.

27 Smith-Bindman R, Lipson J, Marcus R, *et al.* Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. *Arch Intern Med* 2009; **169**: 2078–86.

28 Katzberg RW, Lamba R. Contrast-induced nephropathy after intravenous administration: fact or fiction? *Radiol Clin North Am* 2009; **47**: 789–800, v.

29 Gatto NM, Frucht H, Sundararajan V, Jacobson JS, Grann VR, Neugut AI. Risk of perforation after colonoscopy and sigmoidoscopy: A population-based study. *J Natl Cancer Inst* 2003; **95**: 230–6.

30 O'Sullivan JW, Albasri A, Nicholson B, *et al.* Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open* 2018; **8**: e018557.

31 Glasziou P, Irwig L, Aronson J. Evidence-based medical monitoring: from principles to practice. Oxford (UK): Blackwell Publishing, BMJ books, 2008.

32 Masson S, Latini R, Anand IS, *et al.* Prognostic Value of Changes in N-Terminal Pro-Brain Natriuretic Peptide in Val-HeFT (Valsartan Heart Failure Trial). *J Am Coll Cardiol* 2008; **52**:

997–1003.

33 Wald NJ. Guidance on terminology. *J Med Screen* 2008; **15**: 50–50.

34 Raffle A, Gray J. Screening: Evidence and Practice. Oxford University Press, 2007.

35 UK Government. Population screening programmes. Gov.uk. 2018. https://www.gov.uk/topic/population-screening-programmes.

36 Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M. Ruling a diagnosis in or out with 'SpPIn' and 'SnNOut': a note of caution. *BMJ* 2004; **329**: 209–13.

37 Cecchini M, Lee S, For the Organisation for Economic Co-operation and Development. Tackling Wasteful Spending on Healthcare. 2017 http://networks.sustainablehealthcare.org.uk/sites/default/files/resources/Tackling Wasteful Spending on Health.pdf#page=117.

38 Brownlee S, Chalkidou K, Doust J, *et al.* Evidence for overuse of medical services around the world. *Lancet* 2017; **6736**: 1–13.

39 Glasziou P, Straus SE, Brownlee S, *et al.* Evidence for underuse of effective medical services around the world. *Lancet* 2017; **390**: 169–77.

40 Berwick DM. Avoiding overuse—the next quality frontier. *Lancet* 2017; **6736**: 10–1.

41 Epner PL, Gans JE, Graber ML. When diagnostic testing leads to harm: a new outcomes-based approach for laboratory medicine. *BMJ Qual Saf* 2013; **22 Suppl 2**: ii6-ii10.

42 Gandhi TK, Kachalia A, Thomas EJ, Puopolo AL, Yoon C. Missed and Delayed Diagnoses in the Ambulatory Setting. *Ann Intern Med* 2006; **145**: 488–96.

43 World Health Organization (WHO), The World Bank. Tracking Universal Health Coverage: First Global Monitoring Report. 2015 DOI:978 92 4 156497 7.

44 Berwick D, Hackbarth AD. Eliminating Waste in US Health Care. *JAMA* 2012; **307**: 1513.

45 Saini V, Brownlee S, Elshaug AG, Glasziou P, Heath I. Addressing overuse and underuse around the world. *Lancet* 2017; **6736**: 10–3.

46 Saini V, Garcia-Armesto S, Klemperer D, *et al.* Drivers of poor medical care. *Lancet* 2017; **6736**: 1–13.

47 Brownlee S, Chalkidou K, Doust J, *et al.* Evidence for overuse of medical services around the world. *Lancet* 2017; **6736**: 1–13.

48 Wennberg JE, Fisher ES, Skinner JS. Geography and the debate over Medicare reform. *Health Aff (Millwood)*; : W96-114.

49 NHS Right Care. The 2nd Atlas of Variation in NHS Diagnostic Services in England. 2017.

50 Barros AJD, Santos IS, Matijasevich A, *et al.* Patterns of deliveries in a Brazilian birth cohort: Almost universal cesarean sections for the better-off. *Rev Saude Publica* 2011; **45**: 635–43.

51 Ataklte F, Erqou S, Kaptoge S, Taye B, Echouffo-Tcheugui JB, Kengne AP. Burden of undiagnosed hypertension in sub-saharan africa: A systematic review and meta-analysis. *Hypertension* 2015; **65**: 291–8.

52 Corallo AN, Croxford R, Goodman DC, Bryan EL, Srivastava D, Stukel TA. A systematic review of medical practice variation in OECD countries. *Health Policy (New York)* 2014; **114**: 5–14.

53 Zhi M, Ding EL, Theisen-Toupal J, *et al.* The Landscape of Inappropriate Laboratory Testing: A 15-Year Meta-Analysis. *PLoS One* 2013; **8**: e78962.

54 Gontijo RV, Proietti FA, Santos Amaral CF, De Rezende NA. Appropriateness use of coronary angiography in patients with suspected ischemic heart disease in Brazil. *Int J Cardiol* 2005; **104**: 348–9.

55 Carpeggiani C, Marraccini P, Morales MA, Prediletto R, Landi P, Picano E. Inappropriateness of cardiovascular radiological imaging testing; a tertiary care referral center study. *PLoS One*

2013; **8**: 1–5.

56 Nair R, Aggarwal R, Khanna D. Methods of Formal Consensus in Classification/Diagnostic Criteria and Guideline Development. *Semin Arthritis Rheum* 2011; **41**: 95–105.

57 Lord Carter of Coles. Operational productivity and performance in English NHS acute hospitals: Unwarranted variations. *Dep Heal* 2016; : 87.

58 Busby J, Schroeder K, Woltersdorf W, *et al.* Temporal growth and geographic variation in the use of laboratory tests by NHS general practices: Using routine data to identify research priorities. *Br J Gen Pract* 2013; **63**: 256–66.

59 Chassin MR, Galvin R. The Urgent Need to Improve Health Care Quality Institute of Medicine National Roundtable on Health Care Quality. *JAMA* 1998; **280**: 1000–5.

60 Elshaug AG, Watt AM, Mundy L, Willis CD. Over 150 potentially low-value health care practices: an Australian study. *Med J Aust* 2012; **197**: 556–60.

61 Centers for Disease Control and Prevention, National Center for Health Statistics. National Ambulatory Medical Care Survey: 2012 Summary Tables. 2012; : 5.

62 Koch H, van Bokhoven MA, ter Riet G, *et al.* What makes general practitioners order blood tests for patients with unexplained complaints? A cross-sectional study. *Eur J Gen Pract* 2009; **15**: 22–8.

63 Foot C, Naylor C, Imison C. The quality of GP diagnosis and referral. 2010 http://amapro.isabelhealthcare.com/pdf/Kings_Fund_Diagnosis_and_Referral_2010.pdf.

64 Mafi JN, Parchman M. Low-value care: an intractable global problem with no quick fix. *BMJ Qual Saf* 2018; 27(5):333-336.

65 Achotegui J. The Ulysses Syndrome: The immigrant Syndrome with Chronic and Multiple Stress. 2014.

66 Essex C. Ulysses syndrome. *BMJ.* 2005; 330(7502):1268.

67 Rang M. The Ulysses syndrome. Can Med Assoc J. 1972; 106(2): 122-123

Chapter Three: Identifying tests with the greatest temporal change in use

3.1 Publication statement

A manuscript with components of this chapter has been accepted by the BMJ and is awaiting publication.

3.2 Introduction

The NHS is experiencing an unprecedented rise in spending.[1] Since 2003, net expenditure has increased by over £50bn, an approximate 80% rise.[1] These increasing costs threaten the sustainability of the NHS.

Primary care accounts for 90% of all NHS care,[2] yet few studies have explored primary care activities.[2] There are no official statistics on one of the core services provided by UK General Practitioners (GPs) – the use of tests. Prior research focused on only a few specific tests in a small population sample.[3]

In this chapter, I identify specific tests that have been subject to the greatest changes in use from UK primary care.

3.3 Aim

To identify tests with the greatest temporal change in test use.

3.4 Objectives

1. To identify specific tests that have been subject to the greatest increase in use from a) 2000/1 to 2015/16 and b) 2004/5 to 2015/16.
2. To identify specific tests that have been subject to the greatest decrease in use from a) 2000/1 to 2015/16 and b) 2004/5 to 2015/16.

3.5 Methods

This chapter was conducted in line with the REporting of studies Conducted using Observational Routinely-collected Data (RECORD) checklist (Appendix Chapter 3).[38]

3.5.1 Study population

I obtained electronic health record data for patients registered with general practices contributing to the Clinical Practice Research Datalink (CPRD) during April 1st 2000 to March 31st 2016. The CPRD, a large database of anonymised electronic health records from UK primary care, contains patient-level data covering approximately 7% of the UK population.[5] CPRD data have been validated extensively and are representative of the UK population in terms of age, sex,[5] and ethnic background.[6] I included patients of any age if their records were acceptable for research purposes (a data quality indicator provided by CPRD) and were registered at practices with continuous high-quality data reporting (CPRD defined up-to-standard)[7] at any time during the study period. The protocol was approved by the Independent Scientific Advisory Committee (ISAC) of the MHRA (ISAC protocol number 17_06R – available in Appendix 3). Ethics approval for observational research using the CPRD with approval from ISAC was granted by a National Research Ethics Service committee (Trent MultiResearch Ethics Committee, REC reference number 05/MRE04/87).

3.5.2 Included tests

I selected 44 specific tests (28 laboratory, 11 imaging and five other, miscellaneous tests) to examine. Twenty-five tests were selected based on the following criteria:

1. Specific guidance on their use in primary care is stated in one or more of the following: Quality Outcomes Framework (QOF), National Institute of Health and Clinical Excellence (NICE), Choosing Wisely or NICE Do Not Do guidelines.
2. They were one of the two most frequent laboratory or imaging tests ordered from Oxfordshire primary care (information attained directly from Oxford University Hospital (OUH)).

This list of tests was then discussed and refined with the patient and public involvement (PPI) group that supported Chapters 3 and 4.

A further 19 tests were identified as part of data cleaning and in consultation with GPs as tests that are typically ordered as one test but return many results. For instance, a liver function test returns results for Alanine aminotransferase (ALT), Aspartarte aminotransferase (AST), Alkaline phosphatase (ALP) and Gamma glutamyl transferase (GGT) amongst others, all of which have individual Read codes and are recorded separately. Selection of these additional tests occurred before data analysis but were not identified at the protocol stage.

The CPRD contains primary care data, while secondary care data is captured in a separate database and was not included in this study. We also excluded physical examination findings, vital signs, and body weights.

To avoid double counting, if the same code was recorded multiple times for the same patient on the same day, it was counted as only one test. Similarly, codes likely referring to the same test, or separate components of a single test (e.g. individual components of a full blood count), were grouped and counted as one test.

3.5.3 Statistical analysis

The number of tests ordered was calculated for each year, stratified by age and sex. I calculated total person-years of observation in each age and sex stratum for each year. Patients alive and registered for the entire year contributed 1 person-year of observation to the total. Patients who were born, died, registered, or deregistered during the year were included, but their contribution to the person-year calculation was adjusted proportionately (e.g. a patient who was registered and alive for only 6 months contributed 0.5 person-years).

I calculated crude rates of test use per 10,000 person-years in each age and sex stratum for each year. For comparison across years, I standardised crude rates to the mid-2015 UK population.[8] I calculated standardised rates for each of the 44 specific tests.

Temporal changes in age- and sex-standardised rates were modelled using joinpoint regression.[9] I examined temporal changes from 2000/1 to 2015/16, and, to account for any potential coding

inconsistencies early in our study period, I conducted a sensitivity analysis for the period after the introduction of QOF (2004/5 to 2015/16). The locations of significant changes in slope ('joinpoints') were identified, as were the annual percentage changes (APC) between joinpoints. I also calculated the average annual percentage change (AAPC)[10] across the entire study period (2000/1 to 2015/16) and for the period after the introduction of QOF (2004/5 to 2015/16).

I constructed joinpoint regression models as this method allows an estimation of the rate of test use over a defined time point (e.g. 2000 to 2015) and generates the location (in my case, year) where a significant change in rate has occurred.[11] The location where a significant change in rate occurs is referred to as a "joinpoint". An alternative to joinpoint regression is Poisson regression. The construction of Poisson regression models would have allowed the first, but not latter, to be estimated. To achieve the latter goal, joinpoint models begin with a null hypothesis – there are no significant changes in the rate of test use over time – and compares this against the alternative hypothesis: there are time points where the rate of test use did significantly change. The alternative hypothesis begins with a certain number of prespecified joinpoints and then permutation tests are repeatedly applied until the model of best fit is constructed.[12] Each permutation test has a significant level of 0.05 to avoid type I errors that may occur with multiple tests without individual adjustment. The model of best fit has the lowest sum of residual squares.[9,12] The number of joinpoints prespecified in the first alternative hypothesis model is dependent on the total number of time points across the model. For my model, a maximum of 2 joinpoints was prespecified, as recommended.[9]

Locations of a significant change in temporal slope were identified (a 'joinpoint') and the annual percentage change (APC) of the slope between joinpoints was determined. Thus, our calculated APC represent the annual percentage change for the years between joinpoints. Calculation of the AAPC allows for comparison of tests over the same time interval and for comparison between time intervals. To determine the annual percentage change over the entire study period (2000/1 to 2015/16) and over the post-QOF period (2004/5 to 2015/16) we calculated the AAPC. The AAPC is computed as a weighted average of the APCs within the joinpoint model, with the weights equal to the length of the

respective APC intervals. It is thus valid even if the joinpoint model indicates that there were changes in trend over the interval.[10]

Sarah Stevens extracted the data using Stata (version 14.2) and I performed the statistical analyses using R (version 3.4.1).

3.6 Results

Data from 725 general practices contributed to the CPRD from April 1st 2000 to March 31st 2016. Table 3.1 shows the age and sex adjusted rates of test use for all 44 tests across the study period. Renal function tests were the most frequently ordered test for most of the study period (2002/3 to 2015/16) (Table 3.1). Full blood count (FBC) tests were used most often in 2000/1 and 2001/2 and then were the second most frequently ordered test during the rest of the study (2002/3 to 2015/16). Liver function tests were the third most commonly ordered test from 2001/2 to 2015/16, with urine dipstick third in 2000/1. CT pelvis and MRI knee had the lowest adjusted rates of use (MRI knee from 2000/1 to 2006/7 and CT pelvis for the rest).

Table 3.1 Age and sex adjusted rates of test use per 10,000 person years

Test	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Bone Profile	279.1	367.6	436.7	615.5	781.6	901.3	990.7	1031.8	1104.1	1123.1	1125.2	1189.8	1228.8	1259.1	1266.8	1331.5
Chest x-ray	196.4	217.7	223.4	231.8	246.8	270.1	270.7	278.9	290.2	290.8	300.2	316.2	367.6	349.9	360.3	339.3
Clotting tests	421.9	511.4	571.3	663.0	756.5	873.4	979.1	1073.6	1182.6	1269.3	1351.4	1419.4	1499.2	1614.8	1649.2	1676.7
Colonoscopy	33.9	42.3	49.9	58.1	68.1	77.2	80.2	83.6	91.0	93.1	93.5	96.4	99.1	93.8	90.3	87.9
C-Reactive Protein	91.6	137.2	193.5	260.7	344.2	421.4	470.6	514.6	587.7	635.0	691.9	752.1	788.9	848.4	894.7	924.4
Creatine Kinase	31.6	50.1	70.4	108.8	163.8	201.3	218.3	216.5	207.9	191.2	176.1	134.8	120.7	118.0	106.2	98.1
CT Brain	2.3	3.3	4.7	9.3	13.9	14.1	13.3	12.6	14.5	15.5	15.7	17.5	17.2	17.3	19.3	20.1
CT Pelvis	0.1	0.1	0.1	0.3	0.3	0.3	0.3	0.4	0.4	0.5	1.3	3.7	4.6	5.3	6.4	7.8
DEXA	11.4	18.7	21.7	26.3	27.6	27.2	29.0	31.6	34.7	40.1	38.1	38.8	43.8	48.7	48.8	46.1
Echocardiogram	17.6	25.5	37.5	59.2	77.2	81.9	86.5	88.9	94.1	95.0	93.0	93.7	91.4	96.2	90.2	91.0
Erythrocyte sedimentation rate	501.1	591.4	642.8	752.0	820.4	898.6	916.8	938.4	991.3	1016.8	1015.2	1050.5	1048.1	1059.6	1031.6	1027.2
Female Sex Hormones	68.3	77.4	83.3	95.6	102.6	112.1	114.6	113.3	121.8	123.7	127.1	135.5	130.0	142.2	140.4	140.5
Ferritin	68.6	96.3	126.3	176.8	215.1	267.3	307.8	362.5	437.0	500.2	565.9	636.7	698.2	792.4	864.1	930.6
Folate	64.5	88.0	107.7	144.4	178.6	209.6	235.8	267.9	322.1	381.2	449.5	513.1	586.8	650.7	705.4	780.7
Full Blood Count	1607.7	1920.7	2159.7	2507.4	2775.3	3077.1	3249.1	3383.3	3632.0	3736.9	3897.8	4073.9	4119.7	4304.5	4270.1	4272.1
Glucose	961.9	1240.1	1490.9	1838.3	2046.9	2234.7	2312.0	2325.2	2480.7	2506.5	2524.9	2562.8	2463.9	2373.7	2068.1	1887.4
HbA1c	366.8	434.8	517.5	620.0	713.6	760.8	773.7	794.4	826.9	860.3	908.5	955.7	1100.3	1380.0	1514.1	1711.1
Iron	16.7	20.7	22.6	29.1	37.5	45.1	52.8	61.7	76.3	83.6	97.3	99.3	109.2	120.5	133.9	166.3
Lipids	933.1	1174.2	1425.4	1824.3	2233.8	2401.9	2519.1	2457.5	2555.3	2545.2	2510.8	2549.7	2574.1	2706.0	2457.0	2434.2
Liver Function Tests	977.7	1339.8	1651.3	2146.6	2609.9	2988.9	3229.2	3291.5	3495.9	3558.8	3650.1	3790.1	3844.3	4024.8	3940.4	4002.8
MRI Brain	0.1	0.1	0.1	0.1	0.3	1.1	1.6	4.3	5.9	7.0	8.3	9.6	9.4	10.5	11.5	11.8
MRI Knee	0.0	0.0	0.0	0.0	0.0	0.0	0.1	2.9	4.7	5.9	7.0	7.7	7.7	8.6	8.4	9.1
MRI Lumbar Spine	7.4	9.9	12.4	17.4	24.1	32.5	36.8	41.9	48.1	54.9	57.0	57.5	57.8	60.9	60.3	60.3
Oestradiol	29.9	35.0	37.8	40.2	42.7	46.5	47.6	47.3	50.9	51.0	50.8	52.2	48.7	49.0	49.1	48.0
Pap Smear	61.1	87.8	96.3	92.5	84.1	107.0	117.9	123.6	116.0	105.2	97.0	88.6	75.0	85.6	82.2	82.6

Progesterone	13.0	15.2	16.9	20.1	23.1	26.2	26.8	26.8	28.9	30.3	32.8	35.9	35.3	35.9	35.7	35.7
Prostate Specific Antigen	102.6	128.4	149.7	187.7	213.0	235.3	248.4	264.4	296.6	311.5	316.8	319.8	321.9	352.6	331.5	326.4
Renal Function	1422.1	1826.8	2188.1	2753.8	3216.9	3571.2	3920.4	3988.6	4178.5	4277.1	4366.6	4527.7	4565.6	4799.9	4758.0	4854.6
Spirometry	418.7	453.6	518.3	675.6	880.7	848.8	857.1	832.7	847.8	809.3	805.0	789.2	817.5	781.7	740.9	700.7
Testosterone	15.6	20.5	24.1	28.8	30.3	34.4	35.8	38.8	45.2	53.1	57.9	63.3	65.0	73.3	72.3	72.5
Thyroid Function Tests	862.0	1068.3	1223.2	1523.6	1755.0	1934.2	1995.1	2032.2	2187.6	2248.8	2279.7	2372.2	2389.2	2498.3	2393.1	2424.0
Troponin	1.0	2.5	4.5	9.0	13.6	17.2	18.3	17.8	18.1	18.0	19.8	21.1	19.2	19.5	17.7	17.1
Upper Endoscopy	41.1	52.3	61.1	66.2	68.0	70.1	72.6	74.5	83.1	84.7	88.9	88.2	91.1	91.4	91.2	96.3
Urine Albumin Creatinine ratio	6.5	7.9	6.0	42.1	143.0	179.7	225.4	242.5	276.1	481.1	497.3	544.5	558.9	680.2	675.1	564.4
Urine Albumin	13.1	23.1	46.8	116.5	205.5	239.3	253.7	265.3	282.7	367.6	336.2	337.4	318.8	310.9	225.8	168.4
Urine Dipstick	1023.9	1157.0	1293.5	1445.8	1448.3	1536.5	1674.3	1712.1	1781.3	1936.4	1888.0	1904.7	1841.8	1855.8	1690.9	1548.2
Urine MCS	476.1	552.8	615.9	682.6	712.6	783.2	866.6	899.0	956.5	975.6	949.9	920.8	918.3	962.5	964.8	924.4
Urine non-illicit drugs	9.9	9.5	7.2	7.3	7.5	8.4	8.8	9.5	11.2	15.9	16.9	18.2	19.8	16.4	13.5	12.9
US Pelvis	51.3	58.9	63.4	69.3	75.4	83.1	84.3	91.2	98.4	115.9	124.8	135.5	134.5	136.9	128.5	119.3
Vaginal Swab	133.9	148.5	150.9	129.2	109.7	111.4	106.5	102.2	108.4	105.9	94.7	90.5	79.8	76.0	63.3	51.0
Vitamin B12	75.4	98.4	117.4	157.6	191.6	225.3	252.7	286.6	343.8	401.1	470.1	534.2	609.8	676.9	733.0	811.7
Vitamin D	0.3	0.5	0.9	1.4	2.1	2.8	5.5	8.0	13.8	26.3	52.3	81.7	135.0	191.4	204.2	181.9
X-ray Knee	46.1	49.7	50.8	56.5	61.1	69.1	74.0	79.7	86.3	90.1	93.9	96.8	95.4	100.0	101.1	103.6
X-ray Lumber Spine	56.1	54.0	52.7	54.0	56.9	61.1	60.3	60.3	62.4	59.5	56.8	56.0	54.3	56.2	53.8	54.3

3.6.1 Identifying the test with the greatest increase in use

Table 3.2 lists all 44 tests, ranked by their average annual percentage increase (AAPC). For the period from 2000/1 to 2015/16, 40 of the 44 tests increased in use. MRI knee (69.0% (38.4% to 106.5%)), vitamin D (53.7% (95%CI: 50.2% to 57.3%)) and MRI brain 47.3% (95%CI: 39.5% to 55.5%)) saw the greatest temporal increase in use. For the period from 2004/5 to 2015/16, vitamin D (49.6% (45.9% to 53.3%)), MRI knee (46.5% (27.5% to 68.2%)) and CT Pelvis (37.1% (21.7% to 54.4%)) had the greatest increase in use.

Of the miscellaneous tests, colonoscopy tests saw the greatest increase in use from 2000/1 to 2015/16 (6.1% (95%CI: 5.1% to 7.2%)). Whereas, from 2004/5 to 2015/16, upper gastrointestinal endoscopy saw the greatest increase in use. Table 3.3 shows the three laboratory, imaging and miscellaneous tests that were subject to the greatest increases in use from 2000/1 to 2015/16.Table 3.4 shows the equivalent from 2004/5 to 2015/16.

Table 3.2 Average annual percentage increase in test use

	2000/1 to 2015/16 (95%CI)	Rank %increase 2000/1 to 2015/16	2004/5 to 2015/16 (95%CI): Post QOF	Rank %increase 2004/5 to 2015/16
MRI Knee	69.0% (38.4% to 106.5%)	1	46.5% (27.5% to 68.2%)	2
Vitamin D	53.7% (50.2% to 57.3%)	2	49.6% (45.9% to 53.3%)	1
MRI Brain	47.3% (39.5% to 55.5%)	3	33.8% (28.7% to 39.2%)	4
Urine Albumin Creatinine ratio	44.6% (19.8% to 74.6%)	4	15.6% (8.8% to 22.8%)	5
CT Pelvis	28.4% (16.4% to 41.7%)	5	37.1% (21.7% to 54.4%)	3
Urine Albumin	20.3% (12.3% to 28.7%)	6	-1.0% (-4.8% to 3.0%)	40
Ferritin	19.2% (17.6% to 20.9%)	7	14.4% (13.6% to 15.2%)	7
Troponin	18.8% (16.1% to 21.6%)	8	0.8% (-0.6% to 2.3%)	35
Folate	17.7% (16.5% to 19.0%)	9	14.4% (13.5% to 15.2%)	8
Vitamin B12	17.0% (15.9% to 18.2%)	10	14.0% (13.2% to 14.8%)	9
C-Reactive Protein	16.5% (15.4% to 17.6%)	11	9.4% (8.5% to 10.3%)	10
CT Brain	16.3% (12.4% to 20.3%)	12	3.3% (-0.3% to 7.1%)	26
Iron	16.4% (14.7% to 18.2%)	13	14.7% (12.7% to 16.8%)	6
MRI Lumbar Spine	15.4% (14.3% to 16.6%)	14	9.0% (8.0% to 10.1%)	11
Echocardiogram	11.4% (10.1% to 12.8%)	15	1.4% (0.2% to 2.5%)	33
HbA1c	11.0% (9.5% to 12.4%)	16	8.6% (7.3% to 9.8%)	13
Bone Profile	10.9% (10.1% to 11.7%)	17	5.4% (4.9% to 5.8%)	16
DEXA	10.0% (5.7% to 14.5%)	18	6.0% (5.0% to 7.1%)	15
Testosterone	9.7% (8.2% to 11.2%)	19	8.9% (7.1% to 10.6%)	12
Liver Function Tests	9.5% (8.7% to 10.3%)	20	4.5% (4.1% to 5.0%)	19
Clotting tests	9.3% (8.7% to 9.9%	21	7.6% (6.9% to 8.3%)	14

Renal Function	8.6% (7.4% to 9.9%)	22	4.2% (3.3% to 5.1%)	22
Prostate Specific Antigen	8.1% (6.3% to 9.9%)	23	4.2% (2.6% to 5.8%)	23
Creatine Kinase	7.6% (4.1% to 11.3%)	24	-5.4% (-8.7% to -2.0%)	44
Progesterone	7.1% (6.0% to 8.2%)	25	4.0% (3.0% to 5.0%)	24
Thyroid Function Tests	7.1% (6.1% to 8.1%)	26	2.7% (1.8% to 3.6%)	29
Lipids	6.6% (5.7% to 7.6%)	27	0.2% (-0.8% to 1.1%)	37
Full Blood Count	6.5% (5.8% to 7.2%)	28	4.2% (3.4% to 4.9%)	21
Colonoscopy	6.1% (5.1% to 7.2%)	29	2.3% (1.2% to 3.4%)	31
X-ray Knee	5.9% (5.4% to 6.3%)	30	4.8% (4.3% to 5.3%)	18
Ultrasound Pelvis	5.6% (4.6% to 6.6%)	31	4.5% (3.3% to 5.8%)	20
Upper Gastrointestinal Endoscopy	5.6% (4.1% to 7.0%)	32	3.1% (2.3% to 3.9%)	28
Female Sex Hormones	5.0% (4.2% to 5.8%)	33	3.2% (2.7% to 3.8%)	27
Erythrocyte sedimentation rate	4.7% (3.9% to 5.4%)	34	2.2% (1.4% to 3.0%)	32
Spirometry	4.7% (3.6% to 5.9%)	35	-1.3% (-2.0% to -0.6%)	41
Glucose	4.5% (3.5% to 5.5%)	36	-0.9% (-1.8% to 0%)	39
Urine MCS	4.3% (3.6% to 5.0%)	37	2.5% (1.9% to 3.1%)	30
Chest x-ray	3.9% (3.3% to 4.5%)	38	3.9% (3.3% to 4.5%)	25
Urine Dipstick	3.1% (1.8% to 4.4%)	39	0.8% (-0.2% to 1.9%)	36
Oestradiol	2.9% (1.6% to 4.1%)	40	1.0% (-0.4% to 2.5%)	34
Urine non-illicit drugs	1.4% (-2.3% to 5.2%)	41	4.9% (1.0% to 8.9%)	17
Pap Smear	-0.1% (-2.5% to 2.3%)	42	-2.6% (-4.9% to -0.2%)	42
X-ray Lumbar Spine	-0.3% (-1.7% to 1.1%)	43	-0.6% (-1.4% to 0.2%)	38
Vaginal Swab	-5.2% (-6.3% to -4.1%)	44	-5.2% (-6.3% to -4.1%)	43

Table 3.3 Tests with greatest increase, grouped by type (2000/1 to 2015/16)

Laboratory	Imaging	Other, Miscellaneous
Vitamin D	MRI Knee	Colonoscopy
Urine Albumin Creatinine ratio	MRI Brain	Upper Gastrointestinal Endoscopy
Urine Albumin	CT Pelvis	Spirometry

Table 3.4 Tests with greatest increase, grouped by type (2004/5 to 2015/16)

Laboratory	Imaging	Other, Miscellaneous
Vitamin D	MRI Knee	Upper Gastrointestinal Endoscopy
Urine Albumin Creatinine ratio	CT Pelvis	Colonoscopy
Iron	MRI Brain	Spirometry

3.6.2 Identifying the test with the greatest decrease in use

Only the use of vaginal swab tests fell significantly from 2000/1 to 2015/16 (-5.2% (-6.3% to -4.1%)). Two other tests saw a non-significant decrease in their use (x-ray lumbar spine (-0.3% (-1.7% to 1.1%)) and pap smear (-0.1% (-2.5% to 2.3%)). From 2004/5 to 2015/16, four tests significantly decreased in use (Vaginal swab, Spirometry, Pap smear and Creatine Kinase (Table 3.5)).

Table 3.5 Tests that decreased in use from 2004/5 to 2015/16

Test	Temporal change 2004/5 to 2015/16
Spirometry	-1.3% (-2.0% to -0.6%)
Pap Smear	-2.6% (-4.9% to -0.2%)
Vaginal Swab	-5.2% (-6.3% to -4.1%)
Creatine Kinase	-5.4% (-8.7% to -2.0%)

The rate of use for eight tests did not significantly change from 2004/5 to 2015/16 (urine albumin, troponin, CT brain, lipids, glucose, urine dipstick, oestradiol and x-ray lumbar-spine) (Table 3.6).

Table 3.6 Tests that did not significantly change from 2004/5 to 2015/16

Test	Temporal change 2004/5 to 2015/16
Urine Albumin	-1.0% (-4.8% to 3.0%)
Troponin	0.8% (-0.6% to 2.3%)
CT Brain	3.3% (-0.3% to 7.1%)
Lipid testing	0.2% (-0.8% to 1.1%)
Glucose	-0.9% (-1.8% to 0%)
Urine Dipstick	0.8% (-0.2% to 1.9%)
Oestradiol	1.0% (-0.4% to 2.5%)
X-ray lumbar spine	-0.6% (-1.4% to 0.2%)

3.6.2 Temporal patterns of use

I assessed visually the temporal patterns of use of all 44 tests; five temporal patterns in specific test use emerged (Appendix 3).

- *A consistent linear increase:* chest radiography, vitamin B12, C-reactive protein (CRP), iron, folate, and ferritin
- *An initial rapid increase, followed by a less rapid one:* lumbar spine MRI, pelvic CT, knee MRI, DEXA, MRI brain, female sex hormones, knee radiography, testosterone, clotting, renal function tests, urine albumin–creatinine ratio, bone profile, prostate-specific antigen (PSA), full blood count, and liver function tests)
- *An initial increase followed by a plateau:* echocardiography, thyroid function tests, Erythrocyte Sedimentation Rate (ESR), upper endoscopy, urine microscopy and culture (MCS), progesterone, spirometry, and oestradiol

- *An 'inverted U distribution':* glucose, lipids, troponin, urine dipstick, urine microalbumin, colonoscopy, vitamin D, Papanicolaou smears, creatine kinase, pelvic ultrasonography
- *A fall, an increase and then a fall:* lumbar spine x-ray and urine non-illicit drug test

Two tests (HbA1c and vaginal swab tests) did not adequately fitted any category nor were they similar to each other.

3.7 Discussion

3.7.1 Summary of findings

I have analysed the temporal change in test ordering from UK primary care. From 2000/1 to 2015/16, the use of 40 tests increased significantly. Similarly, from 2004/5 to 2015/16, there was a significant increase in the use of 32 tests; four significantly decreased and eight had no significant changes.

For the period from 2000/1 to 2015/16, MRI knee, vitamin D and MRI brain saw the greatest temporal increase in use. For the period from 2004/5 to 2015/16, vitamin D, MRI knee and CT Pelvis had the greatest increase in use.

3.7.2 Strengths and weaknesses in relation to other literature

There are no official statistics on the use of tests in the UK. A 2013 study[3] assessed temporal changes in 29 laboratory tests ordered from UK primary care from 2005 to 2009 in a random sample of 100,000 patients. My study, in comparison, was more inclusive in its selection of tests – I assessed the changes in the use of 44 tests – nor did I restrict the patient population: I assessed test use in all registered patients over a longer period. I also assessed all types of tests over time (laboratory, imaging, and miscellaneous). Busby et al[3] report that faecal occult blood, CRP and haematinics (haematinic tests were not presented individually) tests had the greatest relative increases from 2005 to 2009. Of the 44 tests I examined, the use of CRP tests had the 11th and 10th greatest increase in use from 2000/1 to 2015/16 and from 2004/5 to 2015/16 respectively. I presented the results of the haematinic tests individually: from 2000/1 to 2015/16 and from 2004/5 to 2015/16, the haematinic

tests had the following ranks of increase: iron tests 13th and 6th, ferritin 7th and 7th, vitamin B12 10th and 9th, and folate 9th and 8th. Faecal occult blood tests were not included in my analysis.

To the best of my knowledge, there are no other studies that have attempted to identify which test has increased the most in use. Nevertheless, there are some reports that have examined test use more broadly: an Australian study reported a 54% increase in laboratory test ordering from 2000/1 to 2007/8,[13] while in the USA, CT and MRI scans increased by 14% and 26% annually from 1997 to 2006.[14]

A potential limitation of my analysis is the inconsistency of coding across the study period. To address this, I only included general practices whose data were quality assured by the CPRD. For General Practices to be included in the CPRD, they must report 'significant clinical events',[3,15] including test ordering. Furthermore, I considered the ordering of a test as either (a) a record indicating that the test had been ordered or (b) evidence of the result of the test (for instance, via a letter or numerical results). Letter correspondence is often incorporated into CPRD, with around 90% compliance.[16,17] I therefore captured tests when the order was not directly documented. To mitigate double counting, if the same test appeared in a patient record twice on one day, these data entries were matched and counted as only one test. Furthermore, a systematic review reports that *all* abnormal test results are recorded accurately in CPRD.[18] A separate study[3] reports that, from the year 2000, the proportion of abnormal results, from the total number of tests ordered, has remained relatively constant in CPRD (28.7% from 2000-2004 and 27.0% from 2005 onwards). These results suggest that the coding of all test results (normal and abnormal results) has been consistent from 2000.

It is also known that the introduction of QOF improved the frequency of data entry in GP records;[19] 90% of GPs were electronically ordering tests from this date.[3,20] Thus, to further investigate any effects of incomplete coding in the early 2000's, I conducted a sensitivity analysis, restricting the study period to after the introduction of QOF.

A further potential limitation of my analysis is the contamination of primary and secondary care data. The CPRD contains primary care data, with additional linkage to secondary care if requested (I did

not attain linked secondary data as it was beyond the aim of this study). Thus, given the quality and validation of CPRD data, it is unlikely, but possible, that some of the test ordering I present reflects tests ordered by secondary care clinicians. Assessment of the temporal trends in two specific tests (Knee MRI and Brain MRI) suggests there is little, if any, secondary care contamination in the data. The provision of direct access MRI has been inconsistent around the UK,[21] but substantial national effort to increase direct access MRIs for GP occurred around 2005/6,[22,23] from 2005/6 my results show a substantial increase in knee and brain MRIs. Furthermore, the rate of Knee, and Brain MRIs from my results are very low before 2005 (the least, and 3rd least utilised tests respectively). These results support the notion that my data is not contaminated with secondary care tests (the access to MRI for secondary care clinicians has been relatively stable since 2000). It should be noted that the other MRI test in our analysis (Lumbar spine) has been available direct access for GPs since the mid to late 90's.[24]

I have not examined the appropriateness of the changes in test use. It is plausible that a substantial proportion of the temporal changes represents overtesting. Overtesting is thought to be prevalent and increasing in health care systems, particularly in developed countries.[25–27] However, it is plausible that some of the temporal change I present is justified. In Chapter 8, I explore how the assessment of temporal change in test use can be part of a wider approach to try and quantify overtesting.

Lastly, I present only a subset of tests available to primary care physicians. It is plausible that there are tests that had greater increases in their use. However, given I selected the most common tests as well as those that are listed in guidelines, this is unlikely. Even if there are tests that had greater temporal increases, it is unlikely these tests are ordered commonly and therefore will probably be of less interest to policy makers, clinicians and patients.

3.7.3 Implications

3.7.3.1 Implications for policy makers

There has been a striking increase in the use of tests from UK primary care. Increasing test use may be a direct response to the increasing number and length of GP consultations,[2] with tests reportedly being used for 'strategic, non-medical reasons', such as reassuring patients and ending consultations.[28] Conversely, greater use of tests is likely to result in more consultations and to contribute to further workload, with more test results to review in non-consultation time. Ordering tests will also beget more testing to follow up abnormal test results, amplified through biological and test variability when multiple investigations are ordered at the same time.

Healthcare systems around the world are struggling to remain financially sustainable.[29] The NHS is attempting to identify and mitigate wasteful spending, to meet calls for £22 billion in efficiency savings.[30] The temporal increase in most tests is unlikely to be sustainable. Policy makers must decide if these temporal increases are appropriate. It is plausible that some of the temporal increase I present, represents unnecessary overtesting, however there may be justified reasons to explain some of the temporal changes. Over the study period, GP access to diagnostic tests expanded,[31] many services were diverted from secondary to primary care,[32] and the QOF was introduced. Furthermore, guidelines may explain the temporal pattern of a few specific tests. The sharp increase in HbA1c use may be attributable to the World Health Organization's 2011 guideline supporting the use of HbA1c for diagnosing diabetes mellitus.[33] Similarly, the 2006 Scottish Intercollegiate Guidelines Network (SIGN) guideline, which recommended a clinical diagnosis of urinary tract infection (UTI),[34] also adopted by NICE, may explain the plateau of urine MCS tests around the same time. For most tests, however, further investigation into the causes and appropriateness of these changes is still required.

3.7.3.2 Implications for researchers

Future research should focus on determining the appropriateness of the temporal changes I have reported. Ideally, this research should use individual patient data to measure appropriateness directly against evidence-based criteria. I discuss an approach to achieve this in chapters 6 and 8.

Furthermore, it is unknown if the changes I present are also occurring in secondary care. An analysis of the temporal changes in secondary care test use would be valuable.

3.7.3.3 Implications for patients

In order to achieve shared decision making, much work is required to empower patients to discuss, agree and disagree with their clinician's suggestions. Although more work is needed to improve patient's knowledge about when a test is appropriate and when it is not, our results can be disseminated to patients. With the knowledge that, for instance, CT pelvic scans have seen substantial increase in their use, patients have objective evidence to support them if they feel inclined to question the appropriateness of the test.

There is also evidence to suggest that patients[35] overestimate the benefits of tests and underestimate their harms. Further education about the balance between benefits and risks is likely to be advantageous.

3.7.3.4 Implications for clinicians

Our results can serve as a basis for clinicians to reflect and audit their own test use. For instance, UK primary care physicians may take note of the tests that have seen the greatest increases in use and audit the appropriateness of their own use of these tests. This may also lead to education of their peers.

There is also evidence to suggest that clinicians[36] overestimate tests benefits and underestimate their harms, which suggests clinicians may need further education regarding the pros and cons of test use.

Furthermore, changes in some of the specific tests may indicate clinician's increasing desire to "not miss" a sinister underlying disease. The four haematinic tests (iron, ferritin, vitamin B12, and folate) followed the same linear pattern, as did CRP. These tests are typically used for patients with non-specific symptoms[28] and may reflect clinician's increased desire to rule out disease in these patients. These changes may also reflect, more broadly, the possible medicolegal consequences of undertesting[37] and the lack of disincentives to overinvestigate.

Lastly, I am involved in delivering OpenPathology.net and OpenImaging.net; open data tools (like OpenPrescribing.net (https://openprescribing.net, EBM Datalab, University of Oxford, 2018)).

OpenPrescribing is a website that reports a variety of NHS General Practice prescribing metrics over time. Metrics include measures such as: the percentage of generic preparations of the drug as a percentage of all drugs prescribed. Prescribing metrics are reported on a national level, for individual Clinical Commissioning Groups (CCGs) and individual General Practices. It displays data from the last 5 years and adds additional data live (as it becomes available). OpenPrescribing also contains a CCG-level estimate for potential cost savings.

Like OpenPresribing, OpenPathology and OpenImaging will provide easy access to various analytic approaches to examine test ordering behaviour in primary care. This tool will continue my work on temporal changes in test ordering on a live interface.

3.7.4 How this chapter fits within wider thesis

My thesis explores different ways to assess primary care test use. In Chapter 8 I discuss an approach to identify potentially inappropriate test use. This approach includes the identification of the tests with the greatest temporal increase in use.

3.8 Conclusion

I have ranked 44 tests from those with the greatest to least temporal increase in use. I found that for the period from 2000/1 to 2015/16, MRI knee, vitamin D and MRI brain saw the greatest temporal increase in use and for the period from 2004/5 to 2015/16, vitamin D, MRI knee and CT Pelvis had the greatest increase in use. Most of the tests I studied increased over the two time periods.

3.9 Chapter Summary

- I used data from the CPRD to identify the tests with the greatest temporal change in use.
- The tests with the greatest to least temporal increase: For the period from 2000/1 to 2015/16, MRI knee, vitamin D and MRI brain saw the greatest temporal increase in use and for the period from 2004/5 to 2015/16, vitamin D, MRI knee and CT Pelvis had the greatest increase in use.

- Only vaginal swab tests significantly decreased in use from 2000/1 to 2015/16, whereas spirometry, pap smear, vaginal swab and creatine kinase significantly decreased in use from 2004/5 to 2015/16.

References

1 NHS Confederation. Key statistics on the NHS - NHS Confederation. Resources. 2016. http://www.nhsconfed.org/resources/key-statistics-on-the-nhs (accessed Aug 19, 2016).

2 Hobbs FDR, Bankhead C, Mukhtar T, *et al.* Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14. *Lancet* 2016; **387**: 2323–30.

3 Busby J, Schroeder K, Woltersdorf W, *et al.* Temporal growth and geographic variation in the use of laboratory tests by NHS general practices: Using routine data to identify research priorities. *Br J Gen Pract* 2013; **63**: 256–66.

4 Lord Carter of Coles. Report of the review of NHS pathology services in England. 2006. http://webarchive.nationalarchives.gov.uk/20130107105354/http:/www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_091984.pdf (accessed Jan 5, 2018).

5 Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015; **44**: 827–36.

6 Mathur R, Bhaskaran K, Chaturvedi N, *et al.* Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *J Public Health (Oxf)* 2014; **36**: 684–92.

7 Williams T, van Staa T, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Ther Adv drug Saf* 2012; **3**: 89–99.

8 Office for National Statistics. Population Estimates for UK, England and Wales, Scotland and Northern Ireland: mid- 2015. 2015. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2015 (accessed Aug 21, 2016).

9 National Cancer Institute. Joinpoint Help Manual, Version 4.5.0.1. 2015. http://surveillance.cancer.gov/joinpoint/Joinpoint_Help_4.2.0.0.pdf (accessed Jan 5, 2018).

10 Clegg L, Hankey B, Tiwari R, Feuer EJ, Edwards B. Estimating average annual per cent change in trend analysis. *Stat Med* 2009; **28**: 3670–3682.

11 Institute National Cancer. Statistical Methodology and Applications Branch, Surveillance Research Program. Joinpoint user's guide 4.5.0.1. .

12 Kim H-J, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joipoint regression with applications to cancer rates. *Stat Med* 2000; **19**: 335–51.

13 Bayram C, Miller G, Valenti L, Britt H. Evidence-practice gap in GP pathology test ordering A comparison of BEACH pathology data and Bettering the Evaluation And Care of Health. 2009. https://www.health.gov.au/internet/main/publishing.nsf/Content/9C300FE48F876E95CA257BF0001ACE0E/$File/Evidence-practice gap in GP pathology test ordering.pdf (accessed Dec 13, 2018).

14 Smith-bindman R, Miglioretti DL, Larson EB. Rising Use Of Diagnostic Medical Imaging In A Large Integrated Health System: The use of imaging has skyrocketed in the past decade, but no one patient population or medical condition is responsible. *Health Aff (Millwood)* 2009; **27**: 1491–502.

15 Medicines and Healthcare Products Regulatory Agency. GPRD recording guidelines for vision users. London, 2004 http://www.vision3.homecall.co.uk/general/GPR Data Recording Guidelines.pdf.

16 Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: A systematic review. *Br J Gen Pract* 2010; **60**: 199–206.

17 Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. *BMJ* 1991; **302**: 766–8.

18 Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: A systematic review. *Br J Clin Pharmacol* 2010; **69**: 4–14.

19 Mannino DM, Thorn D, Swensen A, Holguin F. Prevalence and outcomes of diabetes, hypertension and cardiovascular disease in COPD. *Eur Respir J* 2008; **32**: 962–9.

20 Anonymous. GPRD Users' Group meeting. *Scribe* 2005; **8**: 20.

21 Gough-Palmer AL, Burnett C, Gedroyc WM. Open access to MRI for general practitioners: 12 years' experience at one institution - A retrospective analysis. *Br J Radiol* 2009; **82**: 687–90.

22 Barrett A. Waiting times for scans to decrease, vows Department of Health. *BMJ* 2005; : 331:256.

23 Department of health. Best research for best health: a new national health research strategy. 2006; **6**: 435–7.

24 Chawda S, Watura R, Lloyd DC. Magnetic resonance imaging of the lumbar spine: direct access for general practitioners. *Br J Genera* 1997; **47**: 575–6.

25 Brownlee S, Chalkidou K, Doust J, *et al.* Evidence for overuse of medical services around the world. *Lancet* 2017; **6736**: 1–13.

26 O'Sullivan JW, Albasri A, Nicholson B, *et al.* Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open* 2018; **8**: e018557.

27 O'Sullivan JW, Heneghan C, Perera R, *et al.* Variation in diagnostic test requests and outcomes: a preliminary metric for OpenPathology.net. *Sci Rep* 2018; **8**: 4752.

28 Koch H, van Bokhoven MA, ter Riet G, *et al.* What makes general practitioners order blood tests for patients with unexplained complaints? A cross-sectional study. *Eur J Gen Pract* 2009; **15**: 22–8.

29 Kleinert S, Horton R. From universal health coverage to right care for health. *Lancet* 2017;

390: 101–2.

30 Alderwick H, Robertson R, Appleby J, Dunn P, Maguire D. Better value in the NHS The role of changes in clinical practice. 2015.

31 National Audit Office. Department of Health: Managing high value capital equipment in the NHS in England. 2011. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213151/Radiotherapy-Services-in-England-2012.pdf (accessed Oct 11, 2017).

32 Baird B, Charles A, Honeyman M, Maguire D, Das P. Understanding pressures in general practice. King's Fund. 2016. https://www.kingsfund.org.uk/sites/default/files/field/field_publication_file/Understanding-GP-pressures-Kings-Fund-May-2016.pdf (accessed Dec 12, 2017).

33 World Health Organization. Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. 2011 DOI:WHO/NMH/CHP/CPM/11.1.

34 Scottish Intercollegiate Guidelines Network. Management of suspected bacterial urinary tract infection in adults. 2012.

35 Hoffmann TC, Del Mar C. Patients' Expectations of the Benefits and Harms of Treatments, Screening, and Tests. *JAMA Intern Med* 2015; **175**: 274.

36 Hoffmann TC, Del Mar C. Clinicians' expectations of the benefits and harms of treatments, screening, and tests: A systematic review. *JAMA Intern Med* 2017; **177**: 407–19.

37 Medical Protection Society. Fear of legal action impacting on way GPs practise. 2017. http://www.medicalprotection.org/uk/for-members/press-releases/press-releases/fear-of-legal-action-impacting-on-way-gps-practise (accessed March 8, 2018).

38 Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med

2015;12:e1001885. doi:10.1371/journal.pmed.1001885

Chapter Four: Identifying tests with the greatest between-practice variation

4.1 Publication statement

An advanced draft of a manuscript with components of this chapter has been provisionally accepted at BMC medicine.

4.2 Introduction

Previous research has suggested that when there is strong evidence and a professional consensus that an intervention is effective, there tends to be almost no variation in practice.[1,2] Conversely, variation in healthcare resource use has been used to highlight potential overuse or underuse.[2,3]

Despite its contribution to care and expenditure,[4] there is a paucity of research exploring test use variation between general practices. Expenditure from UK GP test use is thought to be more than £3 billion annually,[4] but only one study has previously explored variation in GP test use.[5] However, this study is outdated, only studied a select group of tests and in a small population.

In chapter 3 I explored the temporal trends in UK primary care test use and identified the tests subject to the greatest changes in use. In this chapter, I set out to identify which tests are subject to the greatest variation in use between UK General Practices.

4.3 Aim

To identify the tests with the greatest between-practice variation.

4.4 Objectives

1. To identify specific tests that are subject to the most between-practice variation in their use.

4.5 Methods

This chapter was conducted in line with the REporting of studies Conducted using Observational Routinely-collected Data (RECORD) checklist (Appendix Chapter 4).[36]

4.5.1 Study population

I obtained electronic health record data from patients registered with general practices contributing to the Clinical Practice Research Datalink (CPRD) during April 1st 2015 to 31st March 2016.

The protocol, which underpins this chapter and the previous chapter (*Chapter Three: Identifying tests with the greatest temporal change in use)* was approved by the Independent Scientific Advisory Committee (ISAC) of the MHRA (ISAC protocol number 17_06R) (the protocol is listed in the Appendix 3).

4.5.2 Included tests

I selected 44 specific tests (28 laboratory, 11 imaging and five other, miscellaneous tests) to examine. Twenty-five tests were selected based on the following criteria:

1. Specific guidance on their use in primary care is stated in one or more of the following guidelines/frameworks: Quality Outcomes Framework (QOF), National Institute of Health and Clinical Excellence (NICE), Choosing Wisely or NICE Do Not Do guidelines.
2. They were one of the two most frequent laboratory or imaging tests ordered from Oxfordshire primary care (information attained directly from Oxford University Hospitals (OUH)).

This list of tests was then discussed and refined with the patient and public involvement (PPI) group that supported Chapters 3 and 4.

A further 19 tests were identified as part of data cleaning and in consultation with GPs as tests that are typically ordered as one test but return many results. For instance, a liver function test returns results for Alanine aminotransferase (ALT), Aspartarte aminotransferase (AST), Alkaline phosphatase (ALP) and Gamma glutamyl transferase (GGT) amongst others, all of which have individual Read codes and are recorded separately. Selection of these additional tests occurred before data analysis.

Like the person-years estimate (*4.5.1 Study population),* I grouped tests into their respective general practices, via their practice identification number. To avoid double counting, if the same Read or entity code was recorded multiple times for the same patient on the same day, it was counted as only one test. Similarly, Read and entity codes likely referring to the same test, or separate components of a single test (e.g. individual components of a full blood count), were grouped and counted as one test.

4.5.3 Statistical analysis

I calculated the total numbers of tests ordered from each general practice from April 1st 2015 to 31st March 2016. I then calculated the total person-years of observation in each age and sex stratum for each of the general practices. Patients alive and registered for the entire year contributed 1 person-year of observation to the total. Patients who were born, died, registered, or deregistered during the year were included, but their contribution to the person-year calculation was adjusted proportionately (e.g. a patient who was registered and alive for only 6 months contributed 0.5 person-years).

For each of the 44 specific tests, I calculated crude rates of test use per person-years for each general practice. I then calculated the standard deviation and mean of each test's crude rate, from which I calculated the co-efficient of variation (CoV) (standard deviation/mean x 100).[9] The CoV is the most appropriate measure of variation, because it considers the standard deviation relative to the mean sample size. The use of CoV facilitates a direct comparison of the variation in use between tests controlling for differences in sample size. The use of CoV facilitates a direct comparison of the variation in use between tests controlling for differences in sample size. It is expressed as a percentage (the ratio of the standard deviation to the mean), with larger percentages reflecting greater variation.

To adjust for differences in patient population between general practices, I constructed a Poisson regression model. The Poisson regression model included three covariates: age, sex and deprivation. The age covariate represents the median age of each general practice, the sex covariate represents the proportion of female patients in each practice and the deprivation covariate represents the practice-level Index of Multiple Deprivation (IMD) deciles. I constructed 44 Poisson models (one for each specific tests), which produced an adjusted number of tests for each general practice. I then calculated

the adjusted rates from this estimate by dividing the adjusted number of test by the person-years for each general practice. The adjusted rates were used to calculate the adjusted coefficient of variation for each test. I ranked tests according to their adjusted coefficient of variation.

To identify the tests that had both a high rate of use and a high CoV I calculated the overall median rate of test use and overall median adjusted CoV. I then classified tests into one of four categories: 1. High variability, low rate, 2. High variability, high rate, 3. Low variability, low rate or 4. Low Variability, High rate. These categories reflect a test's measure in relation to the median value, e.g. high variability, low rate refers to tests with a co-efficient of variation above the median CoV, but a rate of test use below the median rate of test use.

In our ISAC protocol, I stated our intention to perform a secondary outcome of examining the geographical variation in total test ordering between the 12 CPRD regions. Please note that the ISAC protocol underpins research in my DPhil and also wider research within a collaborative, multi-university group. As such, this outcome is not relevant for my DPhil and was never intended for publication within my DPhil.

4.6 Results

Data from 444 general practices contributed to the CPRD from April 1st 2015 to March 31st 2016. In total, 16,496,218 tests were ordered for 4,078,091 patients over 3,311,050 person-years. The median age of patients was 40 (IQR: 21 to 58) and the median percentage of females was 50.6% (IQR: 49.7% to 51.3%). The total number of patients that each practice contributed varied (median: 6,955 (IQR: 4,374 to 9,905) and deprivation differed between included general practice, ranging from 1 to 10 (median: 6 (IQR: 4 to 8)).

4.6.1 Identifying the specific test with the most variation in use

The adjusted CoV varied from 158% (95%CI: 153% to 163%) for non-illicit drug monitoring tests (urine, blood or serum) to 5.6% (95%CI: 5.4% to 5.8%) for testosterone tests. Urine Microalbumin (52% (95%CI: 49.9% to 53.2%)), Pelvic CT (51% (95%CI: 50% to 53%)) and Pap smear (49% (95%CI: 48% to 51%) were the second, third and fourth most variable tests. Drug monitoring, pelvic CT and pap smear tests were also the most variable laboratory, imaging and miscellaneous tests respectively. The median co-efficient of variation was 22.7% (IQR: 14.8% to 31.0%). These measures represent the between-practice variation in test ordering adjusted for age, sex and deprivation differences between practices (table 4.1).

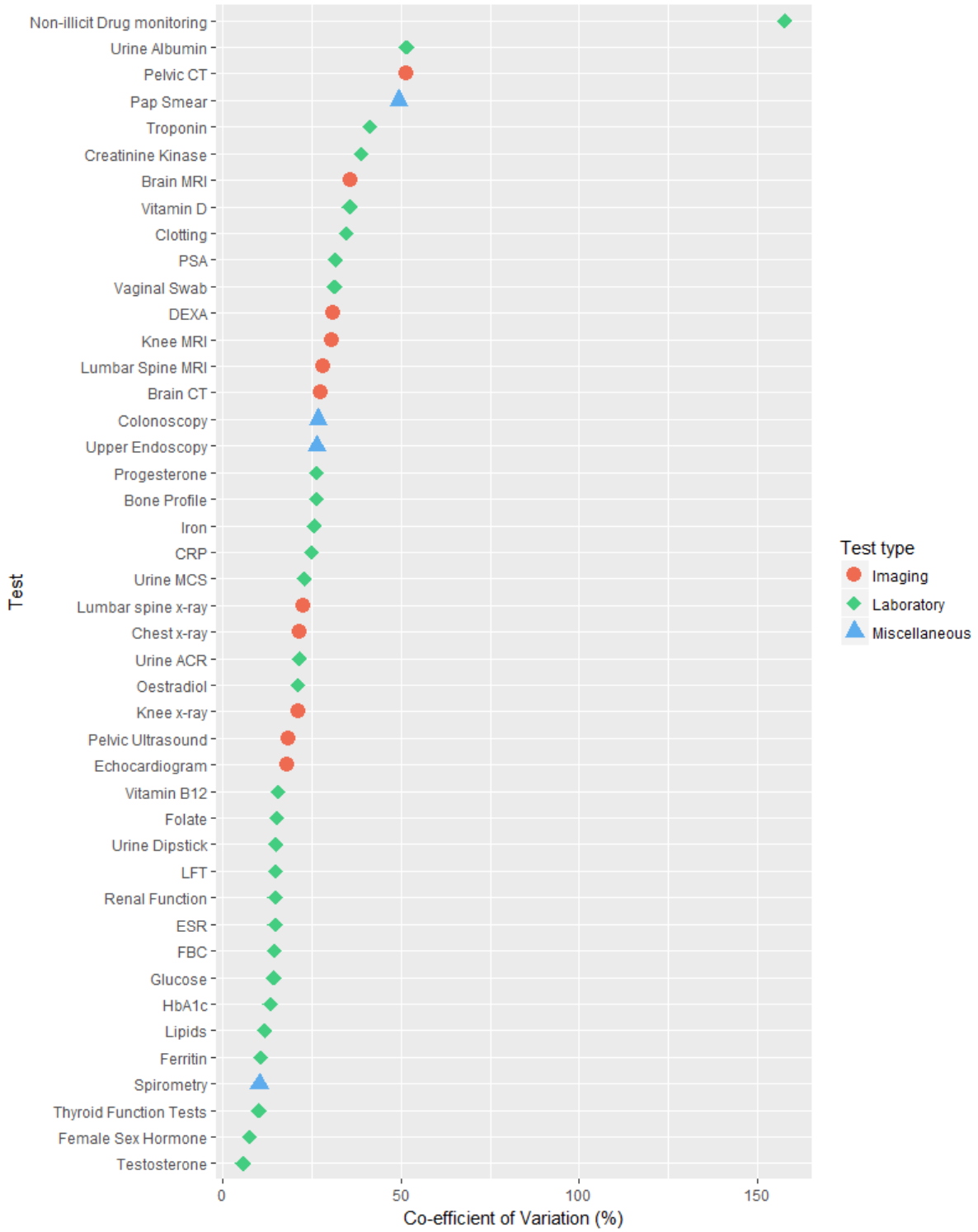


Figure 4.1 Rank order of variability of tests, adjusted for age, sex and deprivation. CT = Computer Tomography, MRI = Magnetic Resonance Imaging, PSA = Prostate Specific Antigen, DEXA = Dual-energy X-ray absorptiometry, CRP = C-reactive Protein, MCS = Microscopy, culture and sensitivities, ACR = Albumin-creatinine ratio, ESR = Erythrocyte sedimentation rate, LFT = Liver Function Tests, FBC = Full Blood Count.

Table 4.1 Adjusted and unadjusted coefficients of variation

Test	Coefficient of variation (adjusted)	Coefficient of variation (unadjusted)	Difference between unadjusted and adjusted co-efficients of variation
Non-illicit Drug monitoring	157.9% (152.8% to 163.1%)	817.9% (813.0% to 822.9%)	660.0%
Urine Albumin	51.6% (49.9% to 53.2%)	177.6% (176.1% to 179.2%)	126.0%
Pelvic CT	51.3% (49.6% to 53.0%)	204.7% (203.1% to 206.3%)	153.4%
Pap Smear	49.3% (47.7% to 50.9%)	225.3% (223.7% to 226.9%)	176.1%
Troponin	41.4% (40.1% to 42.8%)	202% (200.6% to 203.3%)	160.5%
Creatinine Kinase	38.8% (37.5% to 40.1%)	216.6% (215.4% to 217.8%)	177.8%
Brain MRI	35.8% (34.6% to 36.9%)	143.1% (142.0% to 144.2%)	107.3%
Vitamin D	35.6% (34.4% to 36.7%)	132.9% (131.6% to 134.1%)	97.3%
Clotting	34.8% (33.7% to 36.0%)	79.5% (78.2% to 80.7%)	44.6%
PSA	31.5% (30.5% to 32.6%)	59.5% (58.5% to 60.5%)	28.0%
Vaginal Swab	31.4% (30.4% to 32.5%)	163.4% (162.3% to 164.4%)	131.9%
DEXA	30.9% (29.9% to 31.9%)	98.5% (97.5% to 99.4%)	67.5%
Knee MRI	30.7% (29.7% to 31.7%)	165.7% (164.7% to 166.8%)	135.1%

Lumbar Spine MRI	28.2% (27.3% to 29.2%)	118.6% (117.7% to 119.5%)	90.3%
Brain CT	27.4% (26.5% to 28.3%)	140.7% (140.0% to 141.5%)	113.3%
Colonoscopy	26.7% (25.8% to 27.6%)	79.6% (78.9% to 80.4%)	52.9%
Bone Profile	26.3% (25.5% to 27.2%)	73.8% (73.1% to 74.7%)	45.4%
Progesterone	26.3% (25.4% to 27.1%)	139.2% (138.3% to 140.1%)	112.9%
Upper Endoscopy	26.4% (25.5% to 27.2%)	84.7% (83.9% to 85.5%)	58.3%
Iron	25.6% (24.7% to 26.4%)	204.9% (204.1% to 205.9%)	179.4%
CRP	25.0% (24.2% to 25.9%)	59.5% (58.7% to 60.3%)	34.5%
Urine MCS	22.8% (22.0% to 23.5%)	78.4% (77.7% to 79.1%)	55.6%
Lumbar spine x-ray	22.6% (21.9% to 23.4%)	105.5% (104.7% to 106.3%)	82.8%
Chest x-ray	21.6% (20.9% to 22.3%)	71.7% (71.0% to 72.5%)	50.1%
Urine ACR	21.4% (20.7% to 22.1%)	53.3% (52.6% to 53.9%)	31.8%
Knee x-ray	21.1% (20.4% to 21.8%)	71.4% (70.7% to 72.1%)	50.3%
Oestradiol	21.1% (20.4% t o 21.8%)	122.5% (121.8% to 123.2%)	101.4%
Pelvic Ultrasound	18.3% (17.7% to 18.8%)	111.1% (110.5% to 111.7%)	92.9%
Echocardiogram	17.9% (17.4% to 18.5%)	86.3% (85.8% to 86.8%)	68.4%
Vitamin B12	15.5% (15.0% to 16.1%)	71.7% (71.2% to 72.1%)	56.1%

Folate	15.3% (14.8% to 15.8%)	73.5% (73.0% to 74.0%)	58.2%
Urine Dipstick	14.9% (14.4% to 15.4%)	54.7% (54.2% to 55.1%)	39.7%
LFT	14.8% (14.3% to 15.3%)	33% (32.6% to 33.5%)	18.2%
ESR	14.7% (14.2% to 15.2%)	77.9% (77.5% to 78.4%)	63.2%
Renal Function	14.7% (14.2% to 15.2%)	31.2% (30.7% t o 31.6%)	16.4%
FBC	14.6% (14.1% to 15.1%)	32.6% (32.2% to 33.1%)	18.1%
Glucose	14.3% (13.8% to 14.7%)	56.9% (56.4% to 57.3%)	42.6%
HbA1c	13.3% (12.8% to 13.7%)	52.4% (51.9% to 52.9%)	39.2%
Lipids	11.7% (11.3% to 12.1%)	37.9% (37.5% to 38.3%)	26.2%
Ferritin	10.6% (10.2% to 11.1%)	63.6% (63.2% to 63.9%)	52.9%
Spirometry	10.2% (9.9% to 10.5%)	41.3% (40.9% to 41.6%)	31.1%
Thyroid Function Tests	10.1% (9.8% to 10.5%)	33.7% (33.4% to 33.9%)	23.5%
Female Sex Hormone	7.5% (7.3% to 7.8%)	42.3% (42.1% to 42.6%)	34.8%
Testosterone	5.6% (5.4% to 5.8%)	45.8% (45.6% to 46.0%	40.2%

Figure 4.2 shows the adjusted and unadjusted coefficients of variation for the 44 specific tests and Table 4.1 presents the difference between adjusted and unadjusted coefficients of variation for each specific test. Figure 4.3 shows an example of the adjusted and unadjusted rate of test use (for the test CRP), measured against the respective person-years. Similar graphs for the other 43 tests are displayed in appendix 4.

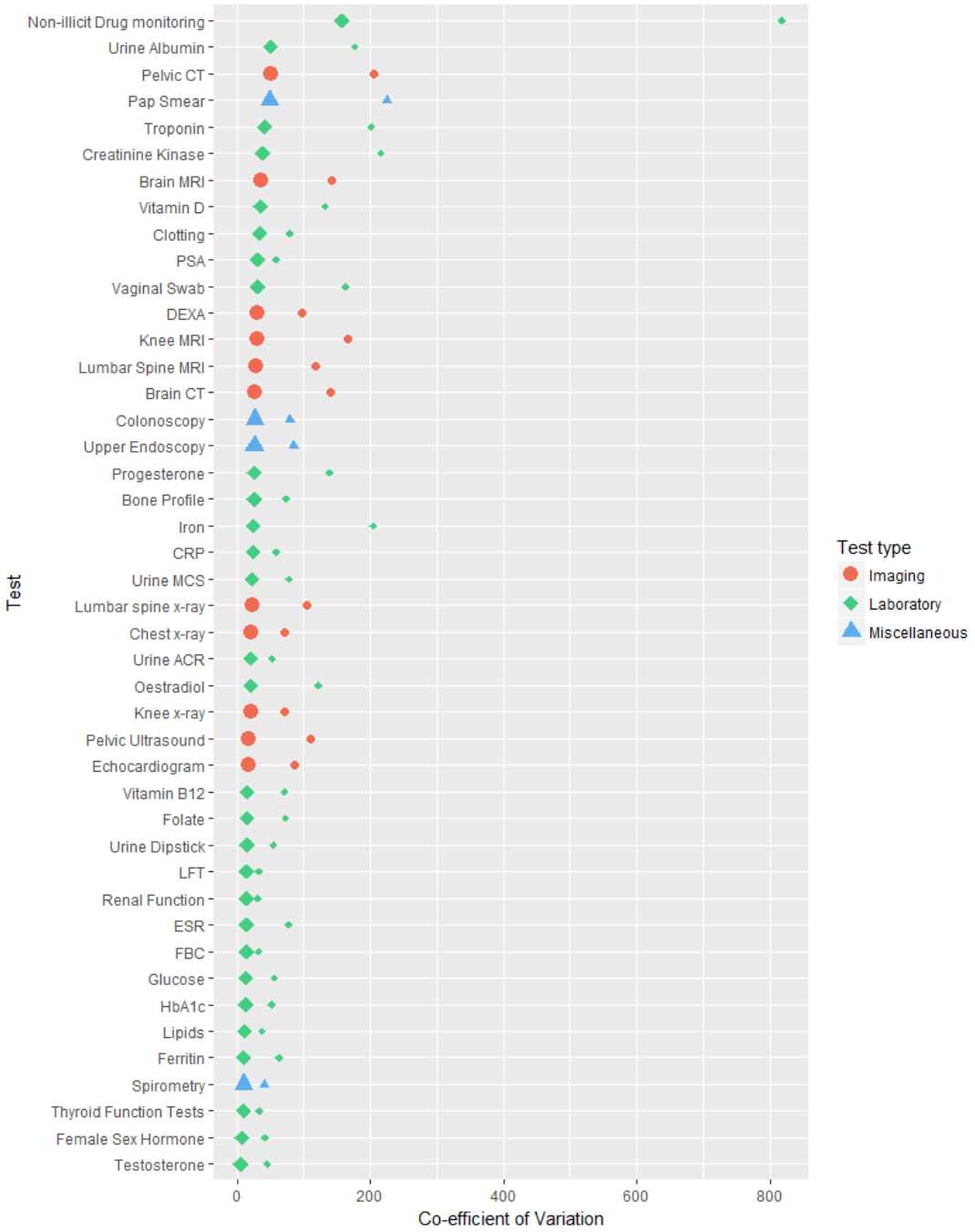


Figure 4.2 Adjusted and unadjusted coefficient of variation:
Larger symbols show adjusted coefficients. CT = Computer Tomography, MRI = Magnetic Resonance Imaging, PSA = Prostate Specific Antigen, DEXA = Dual-energy X-ray absorptiometry, CRP = C-reactive Protein, MCS = Microscopy, culture and sensitivities, ACR = Albumin-creatinine ratio, ESR = Erythrocyte sedimentation rate, LFT = Liver Function Tests, FBC = Full Blood Count.

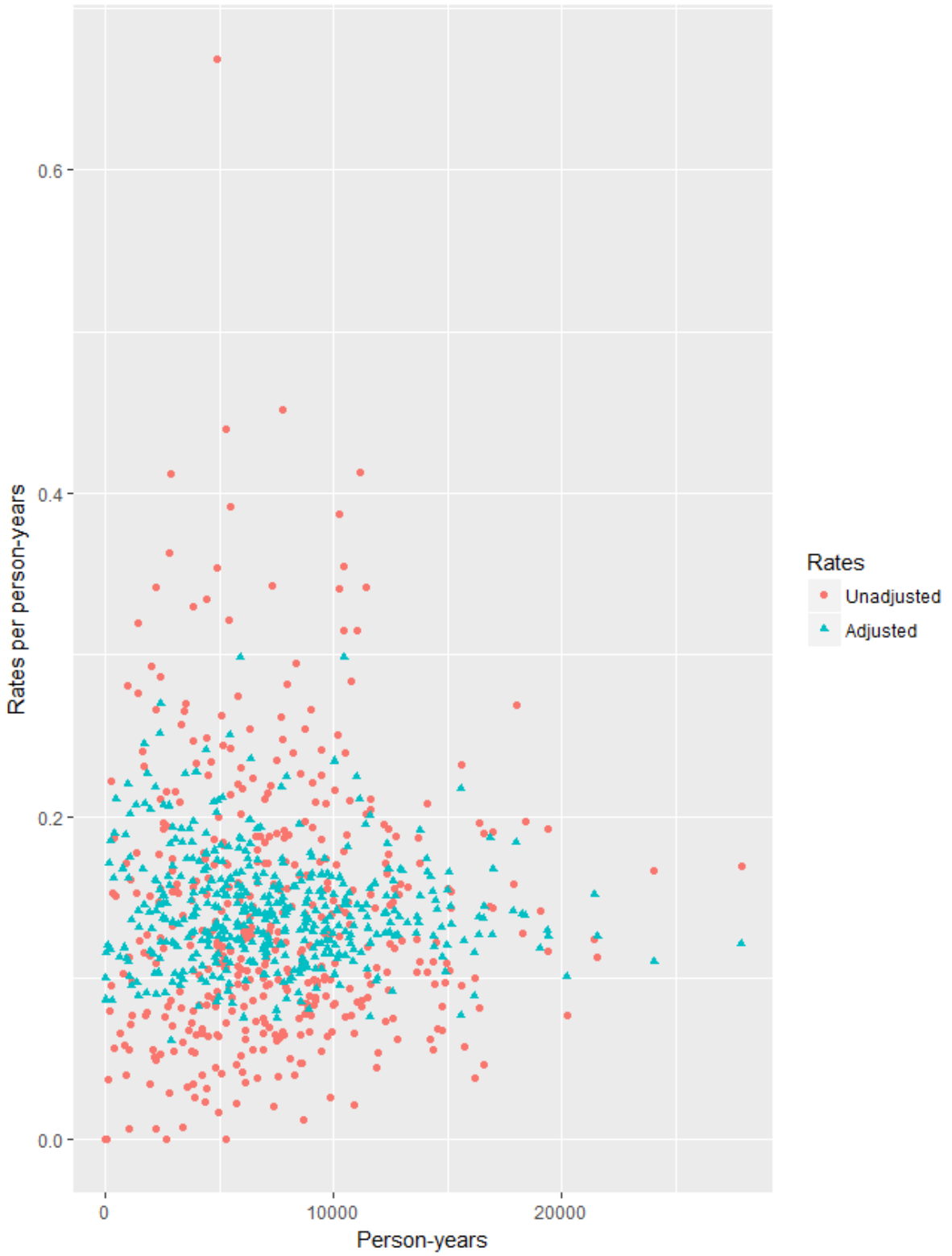


Figure 4.3 Adjusted and unadjusted rates for C-reactive Protein (CRP) use. All 444 practices are represented by both a red and blue data point. Appendix 4 lists the other 43 test graphs like this.

4.6.2 Variability of test use measured against rate of test use

Figure 4.4 plots the adjusted co-efficient of variability of each test against its respective rate. Most tests were classified as high variability, low rate (n = 15, 34%) or low variability, high rate (n = 15, 34%). Seven tests were classified as high variability and high rate (Clotting, Vitamin D, Urine Albumin, Prostate Specific Antigen (PSA), Bone profile, Urine MCS and C-reactive Protein (CRP)). The remaining seven tests were classed as low variability and low rate. Four miscellaneous tests were in the same category (high variability, low rate: colonoscopy, upper endoscopy, pap smear, and vaginal swab), with only one (spirometry) classed as low variability, high rate. Ten of the eleven imaging tests had rates of use below the median; four of these tests were classified as low variability and low rate (Pelvic Ultrasound, lumbar x-ray, knee x-ray and Echocardiogram), while the other six were classified as high variability, low rate (pelvic CT, knee MRI, DEXA, brain MRI, brain CT and lumbar spine MRI). Only one imaging test had an adjusted rate above the median (chest x-ray), which was classified as low variability, high rate. Table 4.2 shows the classification of all tests.

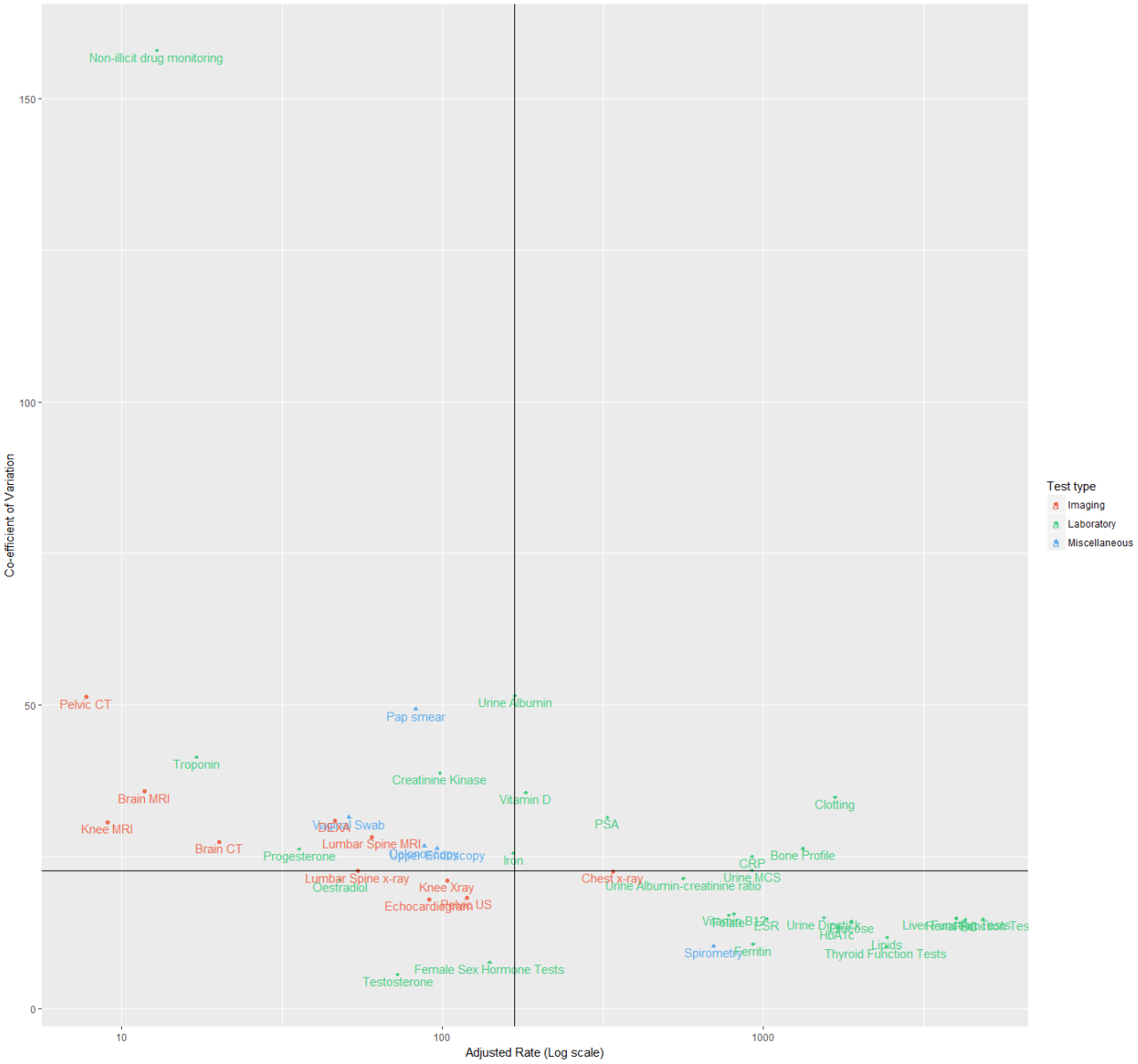


Figure 4.4 Variability and rate of tests. The vertical line represents the median rate of test use and the horizontal line represents the median co-efficient of variation.

Table 4.2 Classification of all 44 tests by rate and CoV

High Variability Low Rate			**High Variability High Rate**		
Test	*Co-efficient of Variation*	*Adjusted rate (per 10,000 person-years)*	*Test*	*Co-efficient of Variation*	*Adjusted rate (per 10,000 person-years)*
Iron	25.6	166.3	Clotting	34.8	1676.7
Creatinine Kinase	38.8	98.2	Bone Profile	26.3	1331.5
Upper Endoscopy	26.4	96.3	C-reactive Protein	25.0	924.4
Colonoscopy	26.7	87.9	Prostate Specific Antigen	31.5	326.5
Pap smear	49.3	82.6	Vitamin D	35.6	182.0
Lumbar Spine MRI	28.2	60.3	Urine Albumin	51.6	168.4
Pelvic CT	51.3	7.8	Urine MCS	22.8	924.4
Vaginal Swab	31.4	51.0			
DEXA	30.9	46.1			
Progesterone	26.3	35.7			
Brain CT	27.4	20.1			
Troponin	41.4	17.1			
Non-illicit drug monitoring	157.9	12.9			
Brain MRI	35.8	11.8			
Knee MRI	30.7	9.1			

Low Variability **Low Rate**			**Low Variability** **High Rate**		
Test	*Co-efficient of Variation*	*Adjusted rate (per 10,000 person-years)*	*Test*	*Co-efficient of Variation*	*Adjusted rate (per 10,000 person-years)*
Female Sex Hormone Tests	7.5	140.5	Renal Function Tests	14.7	4854.6
Pelvic Ultrasound	18.3	119.3	FBC	14.7	4272.1
Knee X-ray	21.1	103.6	Liver Function Tests	13.8	4002.8
Echocardiogram	17.9	91.0	Lipids	11.7	2434.2
Testosterone	5.6	72.5	Thyroid Function Tests	10.1	2424.0
Oestradiol	21.1	48.0	Glucose	14.3	1887.4
Lumbar Spine x-ray	22.6	54.3	HbA1c	13.3	1711.2
			Urine Dipstick	14.9	1548.2
			ESR	14.7	1027.1
			Ferritin	10.6	930.5
			Vitamin B12	15.5	811.7
			Urine Albumin-creatinine ratio	21.4	564.4
			Folate	15.3	780.7
			Spirometry	10.2	700.7
			Chest x-ray	22.6	339.3

4.7 Discussion

4.7.1 Summary of findings

I present a ranking of tests based on the between-practice variation in their use. I analysed over 16 million tests from 444 general practices and ranked tests by their coefficient of variation. The test subject to the greatest variation in use were drug monitoring, followed by urine microalbumin, pelvic CT and then Pap smear. I also identified seven tests with both a rate of ordering and a coefficient of variation above average; Clotting, Vitamin D, Urine Albumin, Prostate Specific Antigen (PSA), Bone profile, Urine MCS and C-reactive Protein (CRP)). These tests are commonly ordered and subject to substantial between-practice variation.

4.7.2 Strengths and weaknesses in relation to other literature

The study of variation in healthcare is not new. Variation in the UK has been documented extensively since the 1930's.[10] While some degree of variation in resource use is expected between healthcare providers, it remains challenging to decipher when variation is justified and when it is not.

Previous research has suggested that when there is strong evidence and a professional consensus that an intervention is effective, there tends to be little variation in practice.[1,2] However, there may also be reasons to explain marked variation. The King's Fund 2011 report on variation within the NHS details some of these reasons.[2] First, the authors report that differences in population characteristics between practices may help explain variation. Specifically, they acknowledge the need for any statistical analysis to account for age, sex and socioeconomic differences between populations; the variation in patient demographics can, in many instances, explain much of the noted variation. My results support this; table 4.1 and figure 4.2 show the difference in CoV before and after demographic adjustment. A strength of my study is its adjustment for demographic differences between practices.

The King's Fund report also notes other factors that could explain variation that persists irrespective of demographic differences. These include: differences in disease prevalence, patient choice, data artefact (differences in data quality), resource availability, local policy and guidelines, and service

configurations. Other previous research also points to the influence of local key-opinion leaders[11] - such as a hospital consultant preferring one test over another - and the variation in management of uncertainty among general practitioners.[12]

Some of these reasons, not all justifiable, may help explain the variation I report. The notable between-practice variation in the use of clotting and drug monitoring tests may reflect regional differences in drug use. In UK primary care, there has been an increase in the use of novel oral anticoagulants (NOAC) (also known as direct acting oral anticoagulants (DOACs);[13] from 2009 to 2015, there was a 17-fold increase in NOAC use.[13] However, there is marked geographical variation in their use.[14] This variation may reflect the non-specific NICE guidance; it states that patients with atrial fibrillation can be anti-coagulated with "apixaban, dabigatran, rivaroxaban or a vitamin K antagonist (VKA)".[15] However, this guidance is now out-of-date compared with recent evidence. A 2017 systematic review and network meta-analysis concluded that "the risk of all-cause mortality was lower with all DOACs" and "several DOACs are of net benefit compared with warfarin".[16] With clear guidance, reflecting the underlying evidence, it is plausible that geographical variation in clotting tests would diminish, especially since standard clotting tests have limited value in 'monitoring' DOAC use.

The high between-practice variation in screening tests is unlikely to be entirely explained by differences in disease prevalence, nor by patient preference. It is surprising that Pap smears, for instance, are subject to much between-practice variation. Given my analysis adjusted for age, sex and deprivation differences between practices, the persistent variation may point to some tests being offered before or after the recommended ages (<25 and >65 respectively) or perhaps at a greater frequency than recommended (every three years for 25-49 year olds and every five years for 50-64 year olds).[17] The noted between-practice variation may also reflect some practices still implementing the original cervical screening guidance. From its start date (1988) to 2003, the NHS Cervical Screening Programme advised to screen women from 20 years old.[18]

A similar phenomenon may be driving the substantial variation in PSA and vitamin D testing. Although neither of these tests are recommended screening tests, both have been subject to sharp

increases in use in UK General Practice (as shown in Chapter 3). In other high-income countries (Canada, and USA), similar increases in vitamin D testing have been noted and deemed largely unnecessary.[19–21] Furthermore, local UK studies have documented the low rate of vitamin D tests returning an abnormal result and the increased number of repeat tests.[22] Similarly, reports from UK GPs suggest that increases in PSA use is unlikely to be appropriate and much test use is likely to be incongruent with the evidence; largely due to Public Health England's mixed messages.[23]

My study is one of the first to use tests themselves, rather than geographical regions, as the unit of analysis. Previous estimates of geographical variation have focused almost exclusively on identifying *regions* (or hospitals) that differed significantly compared with their peers, such as the NHS atlases of variation.[24] One study[5] ranked the variability of test use in UK general practice.[5] However, this study only examined 29 tests, all of which were laboratory tests (not imaging and miscellaneous tests) and only included tests in a restricted sample of patients. Furthermore, this study ranked test variability by their standard deviations. The use of standard deviations to rank the variability of test use is not an appropriate reflection of the variation, as the standard deviation is affected by the rate of testing. Tests that are most commonly ordered are more likely to have higher standard deviations. The use of SD to rank the between-practice variation may make tests that are ordered more commonly appear to have higher between-practice variation. A limitation of our use of CoV is that it may overestimate variation in low ordering tests; to try and mitigate this I presented both the tests with the greatest between-practice variation and the tests with a rate of ordering and a CoV above average. Tests with both a high CoV and high rate of ordering should be the focus of future academic and policy work.

My study is not without limitations. I did not consider the specific indication for test use, and only considered data aggregated at the practice level. My study highlights variation not accounted for demographic differences in practices; my results can be considered population-level, indirect markers of potential waste. I have identified tests that are subject to the greatest between-practice variation; variation that persists despite accounting for demographic differences between practices. In this way, I

have identified tests that would most likely benefit from further investigation into the appropriateness of their use.

My analysis is limited by numerous other factors. I present a large, but not complete subset of tests available to primary care physicians. It is plausible that there are tests which are subject to more variability in their use than the ones I present. However, given I selected the most common tests as well as those that are listed in guidelines, I feel this is unlikely. Even if there are more variable tests they are likely to be ordered uncommonly and thus probably be of less interest to policy makers, clinicians and patients.

4.7.3 Implications

4.7.3.1 Implications for policy makers

Healthcare systems around the world are struggling to remain financially sustainable.[25] The NHS is attempting to identify and mitigate wasteful spending, to meet calls for £22 billion in efficiency savings.[26] My identification of tests subject to the most between-practice variation can assist this process. Variation in healthcare resource use is a recognised indirect marker of inappropriate care, both over and underuse.[27–29] As has been suggested extensively in the literature,[2,5,11,30,31] variation in use is likely to represent professional uncertainty in the appropriate use of a particular resource, in our case, a test. I have identified the tests subject to the most between-practice variation, and also identified the tests with both a high rate of use and high variability of ordering. In this way, I present the tests that are most likely to be subject to over or underuse and could benefit from policy interventions to investigate the appropriateness of their use and, if required, interventions to improve their use.

4.7.3.2 Implications for researchers

Quantification of between-practice variation does not directly assess the appropriateness of test use. In this study, I have not used individual patient data (IPD) nor examined the indications for test use. Therefore, at best, between-practice variation can act as an indicator of inappropriateness, but direct

measures can only be achieved by comparing individual patient data (IPD) with an accepted appropriateness criteria, such as a high-quality evidence-based guideline.[32] My results present a list of 44 tests, ranked from most to least variation. Future research can try and directly quantify the appropriateness of the use of these tests. I describe a method to achieve this in Chapter Six.

Some studies have shown that auditing patient records can directly quantify the proportion of test use that was in line or not in line with a relevant guideline.[32,33] However, this process is not feasible on a national or regional level; our results provide policy makers with an initial step: the identification of tests that are subject to the greatest between-practice variation, which is not accounted for by the demographic differences between practices.

4.7.3.3 Implications for patients

Decision making in healthcare is increasingly shared between clinicians and patients.[34,35] In this patient-centred model, patients need to be empowered to discuss, agree and disagree with their clinician's suggestions. As I have discussed in Chapter 3, my results can help inform patients for their next consultations. For instance, with the dissemination of our results, patients will be aware that chest x-rays are subject to much between-practice variation in use. Patients will thus have objective evidence of variation in the use of the test to support them if they feel inclined to discuss the appropriateness of the test.

4.7.3.4 Implications for clinicians

Currently UK General Practitioners receive no feedback regarding their test use.. The results I present: a ranking of most to least variable tests, can serve as a basis for clinicians to reflect and audit their own test use. For instance, UK primary care physicians may note that CRP tests are both commonly ordered and subject to much between-practice variation. This may lead to refamiliarising themselves with the appropriate indications for this test, which may lead to education of their peers and auditing of their own practice's use.

Furthermore, I am involved with the creation of OpenPathology.net and OpenImaging.net; open data tools (like OpenPrescribing.net) that provides easy access to various analytic approaches that describes test ordering behaviour in primary care. This tool will continue my work exploring between-practice variation in test ordering on a live interface. Open-prescribing.net has received great feedback from clinicians, who have benefited from viewing their prescribing trends in real time.

4.7.4 How this chapter fits within my wider thesis

In my thesis, I use various methods to assess primary care test use. Chapters 3,4 and 5 present population-level, indirect markers of potentially inappropriate test use. In this chapter, I presented one of these metrics: tests with the greatest between-practice variation in use. In Chapter 8 I discuss how my methods can be combined into a step-wise approach to identify potentially unnecessary test use.

4.8 Conclusion

There is wide variation among commonly used tests, which is unlikely to be explained by clinical indication, and since £3 billion annually are spent on tests this represents considerable resource use variation and inefficient management for the NHS. I found that the four most variable tests are non-illicit drug monitoring tests, urine microalbumin, pelvic CT and then Pap smear. I also identified seven tests with both a rate of ordering and a coefficient of variation above average; Clotting, Vitamin D, Urine Albumin, Prostate Specific Antigen (PSA), Urine MCS, Bone profile and C-reactive Protein (CRP)). Tests with the greatest between-practice variation may be worth investigating further to determine whether the noted variation is justified.

4.9 Chapter Summary

- I used data from the CPRD and, after adjustment for practice differences in age, sex and deprivation, quantified the between-practice variation in test use for 44 specific tests. I expressed variation by their coefficient of variation.
- I present the tests in order of most to least variable. I also identify the tests that are subject to above-average rate of use and above-average between-practice variation in use.

- Tests with the greatest between-practice variation may be worth investigating further to determine whether the noted variation is reasonable.

References

1 Wennberg JE. Tracking Medicine. New York: Oxford University Press, 2010.

2 Appleby J, Raleigh V, Frosini F, Bevan G, Gao H, Lyscom T. Variations in health care: The good, the bad and the inexplicable. 2011.

3 NHS Right Care. Diagnostics: The NHS Atlas of Variation in Diagnostic Services. 2012; : 220.

4 Lord Carter of Coles. Report of the review of NHS pathology services in England. 2006. http://webarchive.nationalarchives.gov.uk/20130107105354/http:/www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_091984.pdf (accessed Jan 5, 2018).

5 Busby J, Schroeder K, Woltersdorf W, *et al.* Temporal growth and geographic variation in the use of laboratory tests by NHS general practices: Using routine data to identify research priorities. *Br J Gen Pract* 2013; **63**: 256–66.

6 Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015; **44**: 827–36.

7 Mathur R, Bhaskaran K, Chaturvedi N, *et al.* Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *J Public Health (Oxf)* 2014; **36**: 684–92.

8 Williams T, van Staa T, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Ther Adv drug Saf* 2012; **3**: 89–99.

9 Armitage P, Berry G. Statistical Methods in Medical Research. 1994.

10 Glover JA. The Incidence of Tonsillectomy in School Children. *Proc R Soc Med* 1938; **31**: 1219–36.

11 Wennberg JE. Time to tackle unwarranted variations in practice. *BMJ* 2011; **342**: d1513.

12 Morgan M, Jenkins L, Ridsdale L. Patient pressure for referral for headache: A qualitative study of GP's referral behaviour. *Br J Gen Pract* 2007; **57**: 29–35.

13 Loo SY, Dell'Aniello S, Huiart L, Renoux C. Trends in the prescription of novel oral anticoagulants in UK primary care. *Br J Clin Pharmacol* 2017; **83**: 2096–106.

14 Millett D. NOAC prescribing varies 16-fold between CCG areas. GP Online. 2016. https://www.gponline.com/noac-prescribing-varies-16-fold-ccg-areas/cv-thromboembolic-disorders/atrial-fibrillation/article/1394082.

15 NICE. NICE guideline: Atrial Fibrillation : management. 2014 https://www.nice.org.uk/guidance/cg180/resources/atrial-fibrillation-management-pdf-35109805981381.

16 López-López JA, Sterne JAC, Thom HHZ, *et al.* Oral anticoagulants for prevention of stroke in atrial fibrillation: systematic review, network meta-analysis, and cost effectiveness analysis. *Bmj* 2017; : j5058.

17 NICE Clinical Knowledge Summaries. Cervical screening. NICE. 2017. https://cks.nice.org.uk/cervical-screening#!scenario.

18 Mackie A. What is the right age for cervical screening? Public Heal. Engl. 2014. https://publichealthmatters.blog.gov.uk/2014/04/28/what-is-the-right-age-for-cervical-screening/.

19 Chami N, Simons JE, Sweetman A, Don-Wauchope AC. Rates of inappropriate laboratory test utilization in Ontario. *Clin Biochem* 2017; **50**: 822–7.

20 Morgen EK, Naugler C. Inappropriate repeats of six common tests in a Canadian city: A population cohort study within a Laboratory Informatics framework. *Am J Clin Pathol* 2015; **144**: 704–12.

21 Felcher AH, Gold R, Mosen DM, Stoneburner AB. Decrease in unnecessary vitamin D testing

using clinical decision support tools: Making it harder to do the wrong thing. *J Am Med Informatics Assoc* 2017; **24**: 776–80.

22 Zhao S, Gardner K, Taylor W, Marks E, Goodson N. Vitamin D assessment in primary care: Changing patterns of testing. *London J Prim Care (Abingdon)* 2015; **7**: 15–22.

23 Kole B. Public Health England advice on PSA testing. *BMJ* 2016; **3796**: 3796.

24 NHS Rightcare. The NHS Atlas of Variation in Healthcare. 2015 http://www.rightcare.nhs.uk/atlas/downloads/2909/RC_nhsAtlasFULL_LOW_290915.pdf.

25 Kleinert S, Horton R. From universal health coverage to right care for health. *Lancet* 2017; **390**: 101–2.

26 Alderwick H, Robertson R, Appleby J, Dunn P, Maguire D. Better value in the NHS The role of changes in clinical practice. 2015.

27 Brownlee S, Chalkidou K, Doust J, *et al.* Evidence for overuse of medical services around the world. *Lancet* 2017; **6736**: 1–13.

28 Glasziou P, Straus SE, Brownlee S, *et al.* Evidence for underuse of effective medical services around the world. *Lancet* 2017; **390**: 169–77.

29 Chalmers K, Pearson S, Elshaug AG. Quantifying low-value care: a patient- versus service-centric lens. *BMJ Qual Saf* 2017; **26**: 855–8.

30 Wennberg JE, Wennberg J, Fisher E, *et al.* Unwarranted variations in healthcare delivery: implications for academic medical centres. *BMJ* 2002; **325**: 961–4.

31 NHS Right Care. The 2nd Atlas of Variation in NHS Diagnostic Services in England. 2017.

32 O'Sullivan JW, Albasri A, Nicholson B, *et al.* Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open* 2018; **8**: e018557.

33 Zhi M, Ding EL, Theisen-Toupal J, *et al.* The Landscape of Inappropriate Laboratory Testing:

A 15-Year Meta-Analysis. *PLoS One* 2013; **8**: e78962.

34 Montori VM. Big Science for patient centred care. *Bmj* 2017; **5600**: j5600.

35 Berger ZD, Brito JP, Ospina NS, *et al.* Patient centred diagnosis: sharing diagnostic decisions with patients in clinical practice. *Bmj* 2017; **4218**: j4218.

36 Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015;12:e1001885. doi:10.1371/journal.pmed.1001885

Chapter Five: Variation in diagnostic test requests and outcomes

5.1 Publication statement

A manuscript with components of this chapter has been published in *Scientific Reports*.[1]

5.2 Introduction

The findings of the two previous chapters highlight two recognised indirect markers of waste. Chapter 3 identified which tests have been subject to the greatest temporal increase in use and chapter 4 identified which tests are subject to the greatest between-practice variation in use.

Previous research has explored the variation in test use between regions of the UK. For instance, the NHS atlas of variation in diagnostic services identified regions within the UK that have ordered a greater or lesser number of tests compared to the national average.[2] However, to my knowledge, there is no research that highlights the variation between general practices in terms of the proportion of ordered tests that return an abnormal result (where abnormal refers to a test result outside the reference range).

Given the paucity of metrics available to assess primary care test use, I set out to explore the between-practice variation in the proportion of tests ordered that returned an abnormal result. I chose to focus on C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR) tests because there is evidence to suggest that the ordering of these tests from General Practice can be improved.[3,4] UK General Practitioners (GPs) feel uncertain about the appropriate use of inflammatory marker tests (CRP, ESR) and 'differ in their approach' to ordering[3]. There is also substantial geographical variation in the crude rate of CRP and ESR test requesting.[4]

5.3 Aim

To identify Oxfordshire General Practices that ordered a significantly different proportion of CRP and ESR tests that returned an abnormal result, compared to their peers.

5.4 Objectives

1. To identify Oxfordshire General Practices that ordered a significantly lesser proportion of CRP tests that returned an abnormal result, compared to their peers.
2. To identify Oxfordshire General Practices that ordered a significantly greater proportion of CRP tests that returned an abnormal result, compared to their peers.
3. To identify Oxfordshire General Practices that ordered a significantly lesser proportion of ESR tests that returned an abnormal result, compared to their peers.
4. To identify Oxfordshire General Practices that ordered a significantly greater proportion of ESR tests that returned an abnormal result, compared to their peers.

5.5 Methods

This study was conducted in line with the RECORD (REporting of studies Conducted using Observational Routinely-collected health Data)[6] Statements. The completed checklist is available in appendix 5.

5.5.1 Study design and setting

I retrospectively analysed all CRP and ESR tests requested from Oxfordshire General Practices from 1st January 2016 to 31st December 2016, using data from the clinical laboratories of the Oxford University Hospitals (OUH) Trust. Practices that send most of their laboratory requests to other laboratories or without practice demographic data were excluded.

5.5.2 Data source

This study was conducted as a service evaluation using previously collected, non-identifiable data. The OUH Research and Development Department approved access to the data (study identifier: CSS-BIO-4-4935). I collected routine electronic CRP and ESR test results held by the OUH laboratory, which included the date of the test, anonymised patient ID, the result, and the unique National General Practice Code. We used the National General Practice codes to identify each Practice's 2016 list size, age and gender breakdown and their Indices of Multiple Deprivation (IMD) via Public Health

England[7] and NHS Business Authority (data corresponding to practice demographics relate to Q2 of 2016 (April to June).[8]

The original data file contained quantitative data on the CRP and ESR test results in mg/L and mm/h respectively. I dichotomised CRP results into 'abnormal' (>5 mg/L) or 'normal' (<5 mg/L), and ESR results into 'abnormal' (>20 mm/h for female patients, >14 mm/h for male patients) or normal (<20 and <14 mm/h respectively) on the advice of a Consultant Chemical Pathologist and custodian of OUH laboratory information (Brian Shine) in accordance with the immunoturbidimetric assay used at OUH laboratories (Abbot Architect chemistry analyser (Abbott Diagnostics, Maidenhead, UK)). I calculated the proportion of abnormal CRP and ESR tests ordered for each practice, as "the number of abnormal CRP tests"/"number of CRP tests ordered" and "the number of abnormal ESR tests"/"number of ESR tests ordered").

5.5.3 Statistical analysis

I constructed two separate statistical models exploring the variation in the proportion of abnormal CRP and ESR tests across practices using methods adopted from Spiegelhalter's method for handling over-dispersion of performance indicators.[9] Spiegelhalter suggests using a random-effects model to compare institutional performance,[9] as it controls for 'unmeasured factors that lead to systematic differences between the true underlying rates in institutions'.[9] I therefore used a generalised linear mixed-effects regression model with binomial link function to estimate an 'adjusted average proportion abnormal' of CRP tests and ESR tests for each General Practice. I adjusted for differences in practice populations between practices; namely differences in age, gender and deprivation. I also adjusted for the difference in the proportion of CRP and ESR testing that was 'repeat' testing – tests ordered for patients that received >1 CRP or ESR test in 2016. I adjusted for differences in age and gender by calculating the proportion of a General Practice population >65 years and female. The 2016 practice-specific IMD scores were used to adjust for differences in deprivation. I also used a random effects model to account for other unmeasured differences between practices. These model-predicted adjusted proportions of abnormal CRP and ESR tests for each practice represent practice-level

estimates, accounting for any differences in age, gender, repeat testing and deprivation and other unmeasured differences between practices. Practice list size was ultimately dropped from the model due to co-linearity with the total number of CRP and ESR tests ordered per practice, which was included in the model. Sex was similarly dropped, as it did not improve fit.

I constructed funnel plots to display the variation in the proportion of abnormal CRP and ESR tests between practices. Funnel plots show an 'observed indicator against a measure of its precision, typically the sample size'.[9] I plotted 95% and 99.8% control limits corresponding to two and three standard deviations (SD) from the mean. I also plotted the crude and adjusted proportion of abnormal CRP and ESR tests requested for each General Practice against the total number of CRP and ESR tests ordered per practice. All analyses were performed in R (version 3.4.1).

5.5.3 Data availability

The unadjusted average proportion of abnormal CRP and ESR tests and total tests ordered per practice analysed during the current study are available in appendix 5. The code used to produce the results is published on GitHub (https://github.com/jackosullivanoxford/Scientific-Reports-Paper/blob/master/R_code).

5.6 Results

I analysed 143,745 and 30,758 CRP and ESR tests from 69 General Practices over 2016; on average each practice ordered 2083 CRP and 446 ESR tests annually (table 5.1).

Table 5.1 Practice characteristics

Practice Characteristics	**Number**
Number of General Practices	69
Total number of CRP tests ordered	143,745
Total number of ESR tests ordered	30,758
Median Practice List Size (IQR: Q1 to Q3)	8243 (5342 to 11509)

Median IMD Decile (IQR: Q1 to Q3)	9 (7 to 10)
Mean number of CRP tests ordered per practice annually (SD)	2083 (1044)
Mean number of ESR tests ordered per practice annually (SD)	446 (400)

5.6.1 Variation in the crude proportion of abnormal CRP and ESR tests ordered

Figure 5.1 shows the variation in the adjusted proportion of abnormal CRP and ESR tests requested for each General Practice. The annual proportion of CRP tests that returned an abnormal result varied from 22% to 42%, whereas the proportion of abnormal ESR tests varied substantially more, from 12% to 41%.

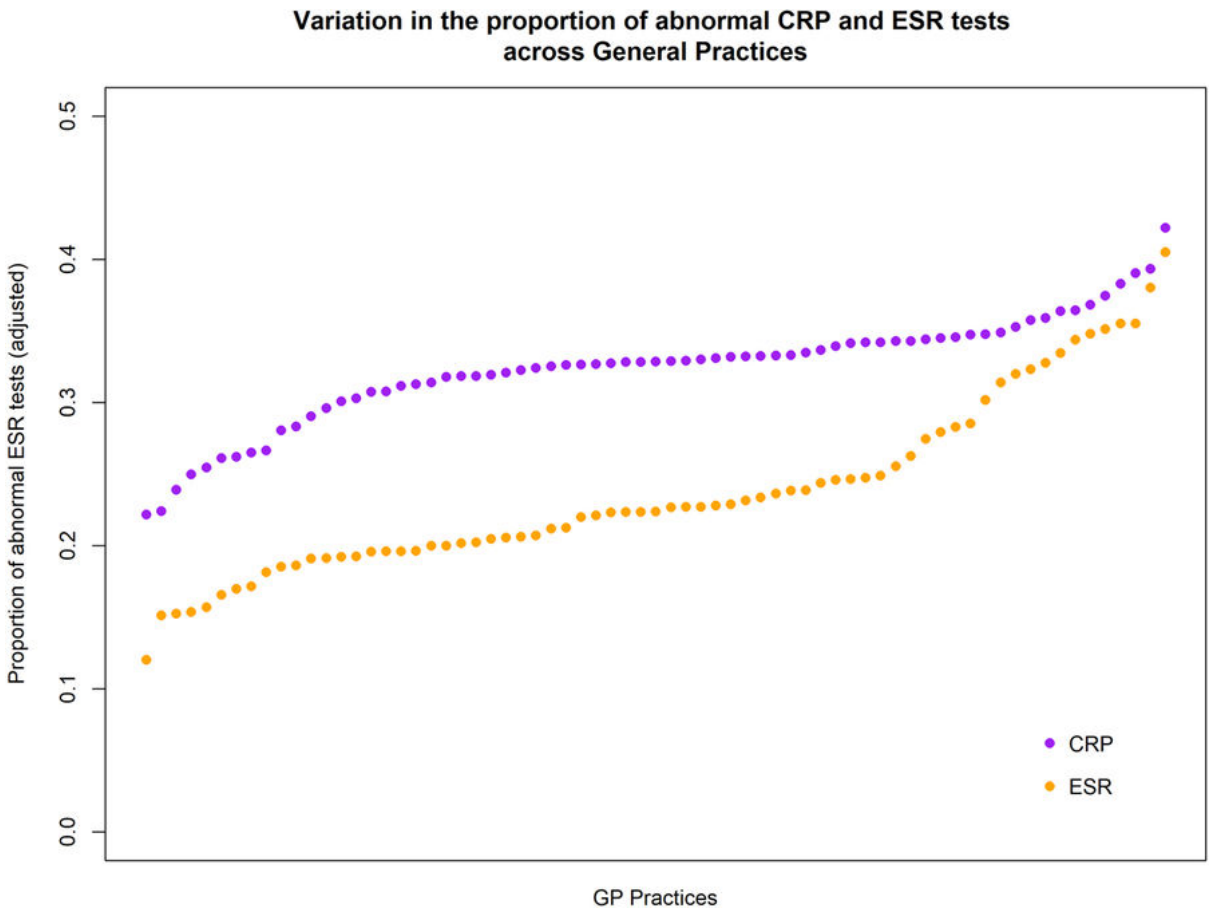


Figure 5.1 Variation in the proportion of abnormal CRP and ESR tests across Oxfordshire General Practices.

5.6.2 Random effects binomial regression model

I report the raw (unadjusted) and adjusted proportions of abnormal CRP and ESR tests for each practice (Table 5.3, in Appendix 5). Figures 2 and 3 show the effect of the binomial regression model on the proportions of abnormal CRP and ESR tests for each practice. Table 5.2 shows the effect of the model on number of practices within the 95% and 99.8% control limits. As expected, the model brings each General Practice's proportion of abnormal CRP and ESR testing towards the mean, and this effect diminishes with increasing sample size.

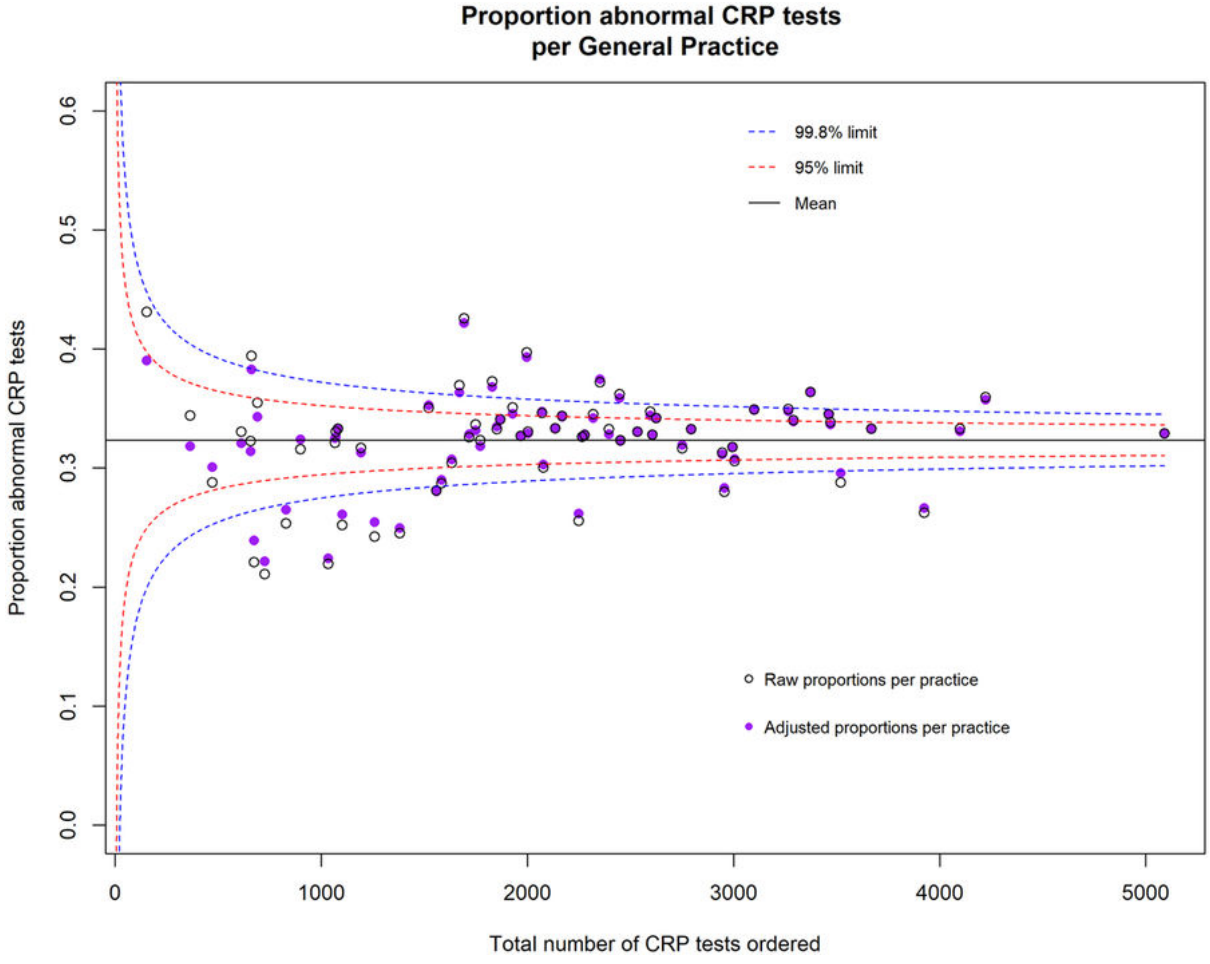


Figure 5.2 Funnel plot comparison of adjusted and unadjusted (raw) average proportion of abnormal CRP tests per GP practice.

The mean proportions of abnormal CRP and ESR tests among all Oxfordshire General Practices were 0.32 and 0.24 respectively. For both the CRP and ESR model, most of the practices (CRP: n = 38; 55%, ESR: n = 43, 62%) were within 2 standard deviations of the mean (Table 5.2). Regarding the CRP model, the proportions of abnormal CRP tests ordered for 20 practices (29%) were greater than 3 standard deviations from the mean, of which 12 practices (17%) were below the 99.8% limit (3 standard deviations). Eleven practices (16%) had proportions greater than 2 but less than 3 standard deviations from the mean, two of these practices (2.9%) were below the 95% limit (2 standard deviations) (Table 5.2).

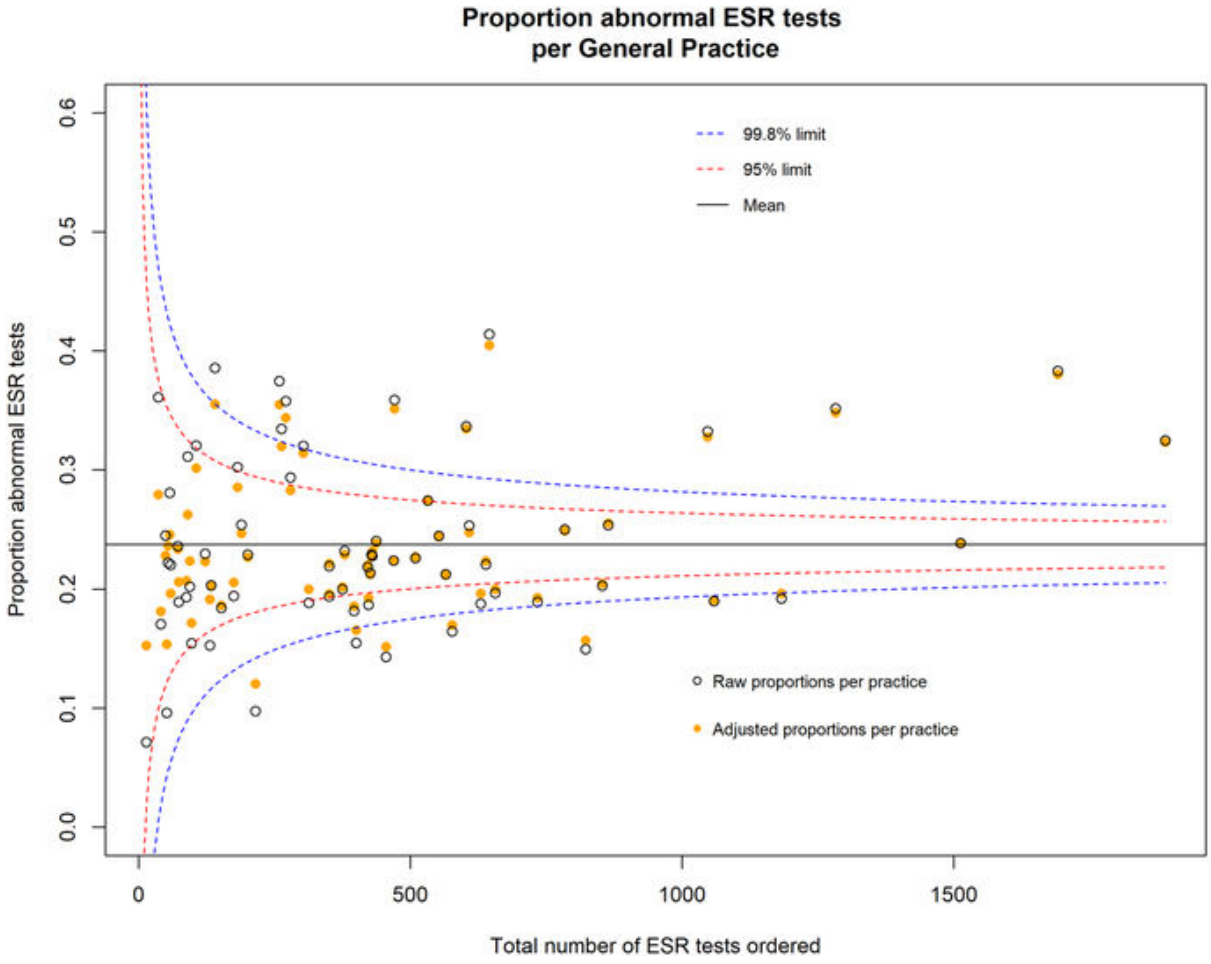


Figure 5.3 Funnel plot comparison of adjusted and unadjusted (raw) average proportion of abnormal ESR tests per GP practice.

Table 5.2 Comparison of adjusted and unadjusted models

Model	Number of practices within 2 SD of mean (%)	Number of practices beyond 2 SD of mean (above or below) but within 3 SDs (%)	Number of practices beyond 3 SD of mean; above or below (%)	Number of practices below 3 SDs from mean
CRP Adjusted	38 (55%)	11 (16%)	20 (29%)	12 (17%)
CRP Unadjusted	33 (48%%)	15 (22%)	21 (30%)	12 (17%)
ESR Adjusted	43 (62%)	10 (15%)	16 (23%)	7 (10%)
ESR Unadjusted	37 (54%)	13 (19%)	19 (28%)	7 (10%)

The proportions of abnormal ESR tests ordered for 16 practices (23%) were greater than 3 standard deviations from the mean, of which 7 practices (10%) were below the 99.8% limit. Ten practices (15%) had proportions greater than 2 but less than 3 standard deviations from the mean, six of these practices (6.7%) were below the 95% limit (Table 5.2).

Two practices (3%) were below the 99.8% control limits for both CRP and ESR ordering, while three practices (4%) were above the 99.8% limit for both CRP and ESR ordering.

5.7 Discussion

5.7.1 Summary of findings

Data were successfully acquired and processed. After adjustment, 12 (17%) and seven (10%) practices had a proportion of abnormal CRP and ESR tests more than 3 standard deviations below the mean. Two practices (3%) were below the 99.8% control limit for both CRP and ESR ordering. Adjustment consisted of controlling for differences in age, gender, deprivation and the total and proportion of repeat CRP and ESR tests between general practices, as well as other unknown differences between the practices in the random effects model.

5.7.2 Strengths and weaknesses in comparison with previous literature

To our knowledge there is no previous work describing the variation in the proportion of abnormal CRP or ESR tests between General Practices, or indeed variation in the proportion of abnormal results for any other test. However, the UK's National Audit Office has developed a similar process for generating the proportion of 'positive' referrals to the 2-Week Wait (2WW) cancer pathway.[10] They monitor and make available the 'conversion rate' of referrals GPs make to specialists under the 2WW pathway.[11] The results are stratified by Clinical Commissioning Group (CCG) and can be compared online.[11]

The use of CRP and ESR in UK General Practice has been examined quantitatively[4] and qualitatively.[3] From 2005 to 2009, CRP use rose 86%, second only to faecal occult blood,[4] ESR use also rose, by 24%. Similarly, requests for CRP and ESR had substantial geographical variation (ESR and CRP were the 11th and 14th most variable lab tests).[4] Qualitative work has shown that UK GPs feel uncertain about when to order CRP and ESR tests and that they differ in how they incorporate the use of these tests into their approach to patients with undifferentiated symptoms.[12]

I have applied robust methods to adjust for known and unknown factors that may account for variations across GP practices, applying an appropriately conservative random-effects model. While it is desirable to account for unknown differences between practices our approach may be excessively conservative, obfuscating GP practices that are truly 2 or 3 standard deviations divergent from the mean.[9] I was unable to identify any further data sources for explanatory variables in our model, such as disease prevalence by small local area, and could therefore only account for differences in age, gender, deprivation, repeat testing and the number of CRP tests requested. There may therefore be other unmeasured and unmeasurable confounders.

A further limitation is the generalisability of our data. I only included data from Oxfordshire General Practices, and thus our results do not reflect UK primary care nationally and may not be consistent across the UK. My methods, however, are still applicable to other regions and could be applied at a national level.

Furthermore, inferences about individual patients from our results would be inappropriate and represent an ecological fallacy. I only suggest our results to be used to identify general practices that would benefit from further investigation; the practice is the unit of analysis, and the unit of interest, not the patient.

Examining CRP and ESR is a strength for many reasons. There is prior evidence that UK GPs are uncertain when and how to use CRP and ESR tests,[3] and there is temporal and geographical variation in the number of CRP and ESR test requests in the UK.[4] Also, unlike radiological or other diagnostic services, CRP and ESR results can be clearly dichotomised into normal and abnormal.

5.7.3 Implications

5.7.3.1 Implications for policy

Ultimately, I envisage that data comparing the proportion of abnormal test results could be used as a feedback tool to give GPs information about their use of tests or for monitoring of GPs at a CCG or national level. Similar feedback is already given to GPs regarding, for example, antibiotic prescribing, with some evidence of impact on prescribing.[13,14] As with all audit and data feedback, such information is best thought of as a measure rather than an indicator, requiring judicious interpretation in the context of population and patient data.[15,16] In this regard, I do not want my results to imply that GPs three standard deviations above or below the mean of their colleagues are ordering CRP and ESR tests inappropriately. Appropriateness of test ordering should be judged using patient level data; my results can only determine GPs that would benefit from patient-level audit. For instance, as pressure to reduce waste continues to increase in the NHS, the Oxfordshire CCG could direct resources to auditing the CRP ordering of the 12 practices with an average proportion of abnormal results below 3 standard deviations, rather than all 69 practices. My methods can be applied to other diagnostic tests and at a national level as an initial step in trying to identify inappropriate use of diagnostic tests.

I suggest that practices below (rather than above) the lower 99.8% control limits should be the focus of audit, particularly practices that order many CRP or ESR tests, as this may reflect excessive or

inappropriate use in patients with a low pre-test probability of an abnormal result. Interpretation of practices with a proportion of abnormal tests above the upper 99.8% control limit is more complex. This could represent GPs with astute history and physical examination skills; or GPs with an excessively high threshold for ordering a CRP or ESR; or GPs that ordered CRP or ESR tests more commonly for patients with a high pre-test probability, for instance using CRP or ESR to monitor chronic disease.

5.7.3.2 Implications for clinicians

My proposed feedback mechanism has implications for clinicians. GPs may be interested to know how their rate of test ordering compares to their colleagues, it may also be of value for GPs to know if they order a lower proportion of tests that return an abnormal result. This feedback may prompt clinicians to reflect on their test use and may lead to positive behaviour change.

I am involved in delivering OpenPrescribing.net: an openly accessible, open source, and open data tool that provides easy access to various analytic approaches to identify outliers for prescribing behaviour in primary care. With new access to data we (I am part of the wider team) are now building the specifications for OpenPathology, taking a similar approach to pathology metrics. Methods developed in this chapter will contribute to some of the metrics displayed in OpenPathology.net. We hope this website will provide feedback on GPs pathology ordering on a national level.

5.7.3.3 Implications for patients

I feel the creation of open data tools (such as OpenPathology,net) will help facilitate conversations between patients and clinicians regarding test use. A normal test is not necessarily an inappropriate test, ruling out disease is an important role of tests. However, the results I present – the proportion of tests that return an abnormal result among general practices – may help patients better understand the two chief functions of diagnostic tests: to rule in or rule out disease. With this understanding, better shared conversations about the roles of a test may occur and consequently facilitate better shared-decision making.

5.7.3.4 Implications for future research

Future research should focus on assessment of the appropriateness of CRP and ESR requests among a range of CRP and ESR requesters (both within 2 SDs of the mean, and those outside 3 standard deviations of the mean), using individual patient data (IPD) either in the form of notes review, GP interview or, if possible, IPD data audit. An assessment of the effect of our data as a GP feedback tool would also be of value, ideally in the context of a randomised controlled trial examining the impact of data feedback on the proportion of abnormal tests ordered.

5.7.4 How this chapter fits within wider thesis

My thesis explores different ways to assess primary care test use. Chapter 3 identifies tests that have been subject to the greatest temporal change in their use, and chapter 4 identifies tests that have been subject to the greatest between-practice variation in their use. This chapter identifies general practices that order a greater proportion of ESR and CRP test that return a normal result. All of these metrics contribute to my broad, step-wise approach to assess primary care test use. This approach is discussed in chapter 8.

5.8 Conclusion

I used robust and conservative methods to identify two practices (3%) that requested a significantly lower proportion of CRP and ESR tests yielding abnormal results. I also identified 12 (17%) and seven (10%) General Practices that requested a significantly lower proportion of CRP and ESR tests yielding abnormal results, respectively. Variation in the proportion of tests with an abnormal result shows promise as an automated tool for auditing variation in care and may contribute to improving quality and cost effectiveness.

References

1 O'Sullivan JW, Heneghan C, Perera R, *et al.* Variation in diagnostic test requests and outcomes: a preliminary metric for OpenPathology.net. *Sci Rep* 2018; **8**: 4752.

2 NHS Right Care. The 2nd Atlas of Variation in NHS Diagnostic Services in England. 2017.

3 Watson J, de Salis I, Hamilton W, Salisbury C, Miles M, Huberman A. 'I'm fishing really' - inflammatory marker testing in primary care: a qualitative study. *Br J Gen Pract* 2016; **66**: e200-6.

4 Busby J, Schroeder K, Woltersdorf W, *et al.* Temporal growth and geographic variation in the use of laboratory tests by NHS general practices: Using routine data to identify research priorities. *Br J Gen Pract* 2013; **63**: 256–66.

5 von Elm E, Altman DG, Egger M, *et al.* Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007; **335**: 806–8.

6 Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015; **12**: e1001885.

7 Public Health England. National General Practice Profiles. 2016.

8 NHS Business Service Authority. General Practice Patient List Size and Demographics. 2016 https://www.nhsbsa.nhs.uk/prescription-data/organisation-data/practice-list-size-and-gp-count-each-practice.

9 Spiegelhalter DJ. Handling over-dispersion of performance indicators. *Qual Saf Heal Care* 2005; **14**: 347–51.

10 National Audit Office. Delivering the cancer reform strategy. London, 2010.

11 Public Health England. Cancer Services. Nancer Cancer Regist. Anal. Serv. 2016.

https://fingertips.phe.org.uk/profile/cancerservices.

12 Hobbs FDR, Bankhead C, Mukhtar T, *et al.* Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14. *Lancet* 2016; **387**: 2323–30.

13 Meeker D, Linder JA, Fox CR, *et al.* Effect of Behavioral Interventions on Inappropriate Antibiotic Prescribing Among Primary Care Practices. *Jama* 2016; **315**: 562.

14 Hallsworth M, Chadborn T, Sallis A, *et al.* Provision of social norm feedback to high prescribers of antibiotics in general practice: A pragmatic national randomised controlled trial. *Lancet* 2016; **387**: 1743–52.

15 Chalmers K, Pearson S, Elshaug AG. Quantifying low-value care: a patient- versus service-centric lens. *BMJ Qual Saf* 2017; **26**: 855–8.

16 O'Sullivan JW, Albasri A, Nicholson B, *et al.* Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open* 2018; **8**: e018557.

Chapter Six: Auditing the use of diagnostic tests against guidelines: A systematic review

6.1 Publication statement

A manuscript with components of this chapter has been published in BMJ Open.[1]

6.2 Introduction

The first three chapters of this thesis report on two recognised indirect markers of potential waste. This chapter directly audits the adherence of diagnostic test use with national or international guidelines.

In this chapter, I set out to quantify the proportion of both underuse and overuse of primary care diagnostic tests. Underuse of tests, failure to order a test when indicated, can lead to diagnostic errors and delays in diagnosis and the delivery of effective treatment, leading to adverse patient outcomes and further healthcare costs.[2,3] Overuse of tests, the delivery of tests with no clear benefit or when potential harms outweigh potential benefits, subjects patients to direct harms, such as radiation exposure, as well as potential adverse outcomes (e.g. contrast nephropathy),[4] incidental findings[5] and overdiagnosis.[6] Overuse is also a waste of finite healthcare expenditure, diverting resources from beneficial tests and treatments.[7–9]

Many drivers encourage inappropriate underuse and overuse of diagnostic tests in primary care. Greater access to tests,[10] the medicolegal consequences of undertesting,[11] few if any disincentives to overinvestigate[7] and clinical performance measures[12] may all contribute to overuse. Increasing primary care workload,[13] time constraints[12] and difficulty keeping up-to-date with rapidly increasing evidence[14] may contribute to both inappropriate underuse and overuse.

Guidelines set the standard of care across most healthcare settings.[15,16] Furthermore, they provide a medicolegal framework,[17] inform healthcare policy and improve both care outcomes and processes of care.[18] Despite some recognised limitations, including varying quality of guidelines,[19–21] guidelines

are often used as markers of healthcare appropriateness.[22–24] Zhi *et al*,[22] for instance, used guidelines as a measure of appropriateness to estimate underuse and overuse of laboratory testing. They estimated that 45% (95% CI 34% to 56%) of secondary care laboratory testing is underused and 21% (95% CI 16% to 25%) is overused.

Despite the increasing use of healthcare resources, rising healthcare expenditure,[25–27] increasing demands placed on primary care[13] and the apparent drivers of inappropriate testing,[7,10–14,28] it is not clear how often diagnostic tests are inappropriately overused or underused in primary care. I therefore conducted a systematic review to quantify the frequency of inappropriate ordering of all types of diagnostic tests from primary care in relation to their respective guidelines and identify tests that are frequently overused and underused.

6.3 Aim

To identify how often diagnostic tests ordered from primary care are ordered, or not ordered, in line with guidelines.

6.4 Objectives

2. To identify which diagnostic tests are most frequently overuse and to what degree.
3. To identify which diagnostic tests are most frequently underused and to what degree.

6.5 Methods

This study was conducted and is reported in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).[29] The completed checklists are available in appendix 6.1.

6.5.1 Protocol and registration

The protocol has been published and is available online (open access) via the International Prospective Register for Systematic Reviews database (registration ID: CRD42016048832).

6.5.2 Search Strategy

I searched Embase (OvidSP) and MEDLINE (OvidSP) databases from January 1999 to October 2017 for studies of any design measuring how often diagnostic test guidelines were followed in primary care (the search strategy is available in appendix 6.2). My search strategy can be summarised as: 'Ambulatory Care AND adherence AND guideline AND diagnostic tests AND inappropriate'. Conference abstracts published after 2015 were also searched for in these databases to capture data not yet published. I also searched the WHO International Clinical Trials Registry Platform (http://apps.who.int/trialsearch/), ClinicalTrials.gov, and the reference lists of included studies.

6.5.3 Eligibility criteria

I included studies of any design if they measured the rate of inappropriate ordering (overuse) or not ordering (underuse) of diagnostic tests ordered from primary care against national or international guidelines. I considered all diagnostic tests ordered in adults. I also included studies that measured diagnostic tests ordered from primary care but performed in secondary care (e.g. upper endoscopy). I included the control arms of randomised controlled trials (RCTs) if they offered exclusively usual care and the preintervention periods of studies that used interrupted time series designs (before and after studies).

I excluded studies if they met the following criteria: >20% of participants were children (>20% under 18 years old); diagnostic tests not ordered by general practitioners; and screening or monitoring tests or publication before 1999 (studies after 1999 were considered to ensure that results would more closely reflect current practice). I defined a screening test as a test on an asymptomatic or symptomatic person without signs or symptoms related to that test.[30,31] I defined monitoring tests as 'a test for a patient with an established diagnosis, for which the test is used to measure progression of the disease'.[32] I excluded studies if they did not give a measure of appropriateness or if appropriateness was measured against local guidelines, such as a guideline specific to a hospital or region, rather than international or national guidelines.

6.5.4 Study selection and data extraction

Three reviewers (myself and Ali Albasri or Brain Nicholson) independently screened titles, abstracts and full texts for eligibility. The same reviewers assessed risk of bias and extracted the following data from included studies: patient demographics, eligibility criteria, name and type of diagnostic test, duration of study (days), guideline name and recommendation, total number of tests performed and the number of tests ordered when the specific guideline recommended not ordering (overuse) or the number of tests not ordered when the guideline recommended ordering it (underuse). The last two data points (overuse and underuse) represent 'measures of inappropriateness'. When studies measured inappropriateness of multiple tests, we extracted data on each test and presented them as individual measures of inappropriateness. When studies measured tests across different periods, we extracted measures for each time point and considered each one as an individual measure of inappropriateness.

The same reviewers assessed the quality of included studies using a modified version of the Hoy risk of bias tool.[33] This tool has been validated to assess the internal and external validity of prevalence studies.[33] My modified version of this tool kept the same domains but adjusted the wording of the tool to reflect prevalence of inappropriate testing rather than prevalence of disease. My tool (and results) is available in appendix 6.3.

6.5.4 Statistical analysis

The primary outcome was the prevalence of inappropriate diagnostic testing. Inappropriate testing was measured in two ways:

1. Overuse: a diagnostic test was ordered when the relevant guideline recommends not ordering it, for instance, imaging for non-red flag low back pain (LBP).
2. Underuse: a diagnostic test was not ordered when the relevant guideline recommended ordering it, for instance, spirometry to confirm or refute the diagnosis of chronic obstructive pulmonary disease (COPD).

I express measures of inappropriateness as percentages (%), where the numerator represents the total number of times a guideline recommendation was not followed, and the denominator represents the

total number of times a guideline recommendation could have been followed. For instance, the number of times imaging was inappropriately ordered for non-red flag headache as a percentage of the total number of patients who presented with non-red flag headache. As our included data are percentages, I calculated Clopper-Pearson 95% confidence intervals (CIs) for each individual measure of appropriateness. I conducted sensitivity analyses with high risk of bias studies excluded.

Where the same guideline and recommendation were used by multiple studies (e.g. five studies measured inappropriate underuse of spirometry testing in patients with COPD[34–38] using the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guideline, I pooled the measures and assessed heterogeneity. I combined measures of inappropriateness using a random-effects meta-analysis with 95% CIs (Clopper-Pearson). I performed double arcsine transformation on prevalence data to stabilise the variance[39] and pooled the data using the inverse variance method.[40] We assessed heterogeneity using the I^2 statistic.[41] I did not combine measures of overuse and underuse, as they have different denominators: overuse involves the total number of tests ordered, whereas underuse involves the total number of times a test should have been ordered. All analyses were performed in R (version 3.4.1).

6.6 Results

6.6.1 Study selection and characteristics

I included 63 studies from 14,716 references identified from independent searches by two authors (JWO and AA or BDN) (figure 6.1). Of the 63 included studies, 55 were observational studies, six were before-and-after studies and two were RCTs. The two RCTs investigated the effect of implementing an intervention to reduce inappropriate testing. These studies were conducted in 15 countries and included 357 171 patients (see table 6.1, in appendix 6). Table 6.2 (in appendix 6) shows the 103 measures of inappropriateness extracted from included studies for 47 different diagnostic tests measured against 77 guideline recommendations (41 measured underuse and 62 measured overuse). Guideline recommendations came from 42 different guideline organisations from 15 countries.

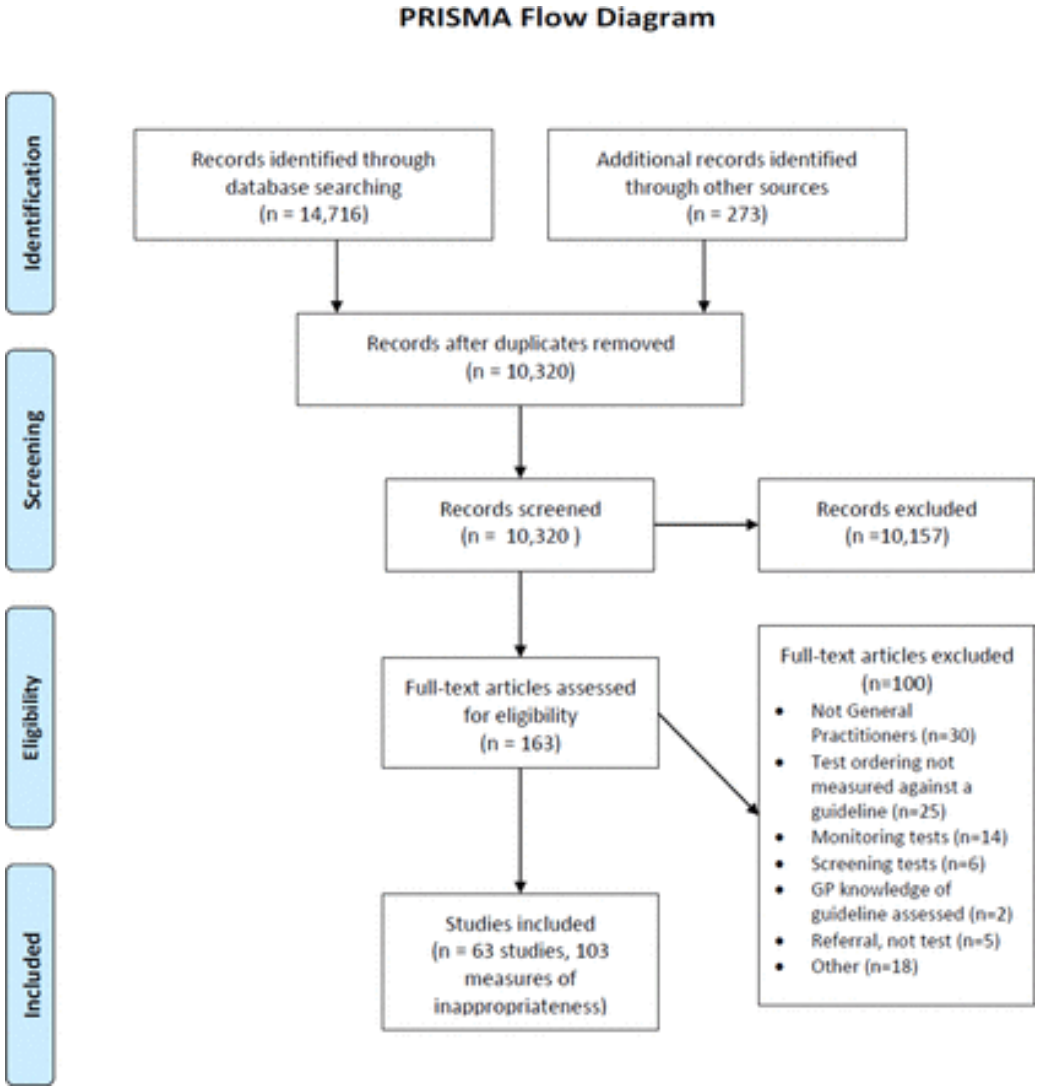


Figure 6.1 PRISMA flow diagram

Fourteen studies measured inappropriateness of more than one diagnostic tests for the same condition (e.g. chest X-ray, electrocardiography and transthoracic echocardiography to confirm or refute a diagnosis of heart failure). Two studies[42,43] measured inappropriateness across multiple time periods. No studies measured both underuse and overuse of the same test.

Included studies measured inappropriateness in one of three ways:

1. Patients with specific symptoms were assessed (prospectively or retrospectively) to see if they had received an inappropriate diagnostic test (overuse) or had not received the appropriate

diagnostic test (underuse) in line with the relevant guideline recommendation (e.g. records for patients with non-red flag LBP to see if they received imaging[44]). Eighteen studies used this method.

2. Patients who had undergone a diagnostic test were identified (via hospital or national databases), and an assessment of whether the test was inappropriate (as per the defined guideline recommendations) via individual patient data was made (overuse). For instance, patients who had an upper endoscopy.[45] Twenty-two studies used this method.
3. Patients with a diagnosis were identified via hospital or national databases and assessed to see whether they had received the appropriate diagnostic test (as per the defined guideline) to confirm or refute the diagnosis via individual patient data (underuse). For instance, assessing if patients with a diagnosis of COPD had spirometry to confirm or refute the diagnosis.[34] Twenty-three studies used this method.

6.6.2 Risk of bias

Two-thirds of the studies (n=44; 70%) were graded as being at low risk of bias, 15 (24%) at moderate risk and 4 (6%) at high risk (Risk of bias table in appendix 6). Moderate or high-risk studies were at an increased risk of non-response bias (>20%), non-objective collection of data and/or unclear intervals between symptom onset and diagnostic test use.

6.6.3 Percentage of diagnostic tests ordered in line with specific guideline recommendations

There was large variation in the rate of inappropriate diagnostic test ordering. The 103 diagnostic test guideline recommendations were not followed 0.2%–100% of the time (table 1, appendix 6); wide variation was largely sustained (0.2%–99.94%) when a further analysis was conducted excluding studies judged to be of high risk of bias. The prevalence of underuse varied 8.2%–100%, whereas overuse varied between 0.2% and 94.2%. Similarly, this variation was essentially maintained on exclusion of high risk studies (under use 9.8%–99.9%, overuse 0.2%–94.2%).

6.6.4 Underused tests

Table 1 (appendix 6) shows that 17 tests were underused more than 50% of the time. Echocardiography was the most frequently studied (n=4, twice in the UK and once each in Poland and Brazil). In patients with heart failure, echocardiography was underused between 54% and 89% (n=3) of the time and in atrial fibrillation 56% (n=1).

For some tests, there was large variation in the rate of underuse (figure 6.2). Underuse of pulmonary function tests (PFTs) to confirm or refute COPD, measured against the GOLD, National Institute for Health and Care Excellence (NICE) and Danish National Board of Health guidelines, varied from 26% to 78% (n=8). None of the studies that studied echocardiography or PFTs were considered high risk of bias and thus results did not change on further analysis excluding high-risk studies.



Figure 6.2 Underuse of tests. ACC, American College of Cardiology; AFib, atrial fibrillation; AHA, American Heart Association; CDC, Centers for Disease Control and Prevention; CXR, chest X-ray; ESC, European Society of Cardiology; FBC, full blood count; FNA, fine needle aspiration; GOLD, Global Initiative for Chronic Obstructive Lung Disease; NICE, National Institute for Health and Care Excellence; PFTs, pulmonary function tests; TB, tuberculosis; TSH, thyroid stimulating hormone; UTI, urinary tract infection.

6.6.5 Overused tests

Eleven tests were overused more than 50% of the time (figure 6.3). Echocardiography was consistently overused, for instance in 'routine perioperative evaluation of ventricular function with no symptoms or signs of cardiovascular disease', whereas other tests (urinary cultures, upper endoscopy and colonoscopy) were overused at varying rates. The overuse of echocardiography was studied in the UK[46] and the Netherlands.[47] The rates of overuse varied between the two settings: between 77% (Netherlands) and 92% (UK). The overuse of urinary cultures for uncomplicated urinary tract infections was studied in the USA,[48,49] Spain[50] and Sweden.[51] The rate of overuse varied from 57% to 77% in the USA, compared with approximately 50% in Sweden and 36% in Spain. Overuse of upper endoscopy was studied widely (n=11) in Australia,[52,53] Saudi Arabia,[54,55] UK,[56] Italy,[57–59] USA[45,60] and Malaysia.[61] The overuse varied markedly, from 7.5% to 54% (n=11), respectively (figure 6.3). Similarly, the inappropriate overuse of colonoscopy varied substantially from 8% in Australia[53] to 52% in Malaysia.[62] None of the above studies were considered high risk of bias and thus results did not change on further analysis excluding high-risk studies.

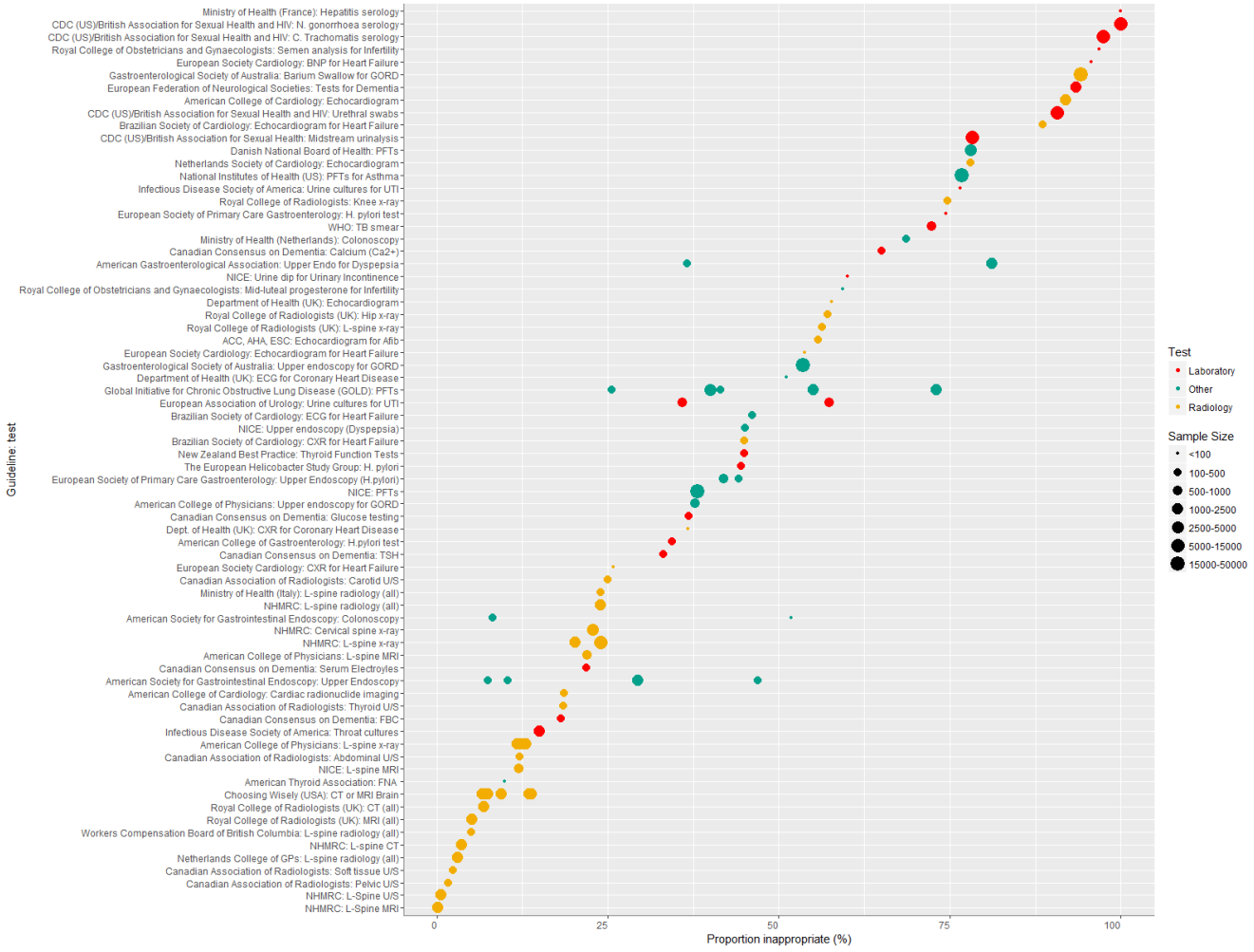


Figure 6.3 Overuse of tests. GORD, gastro-oesophageal reflux disease; GP, general practitioner; LBP, low back pain; NHMRC, National Health and Medical Research Council; NICE, National Institute for Health and Care Excellence; U/S, ultrasound; UTI, urinary tract infection

My results also suggest that the inappropriate overuse of CT and MRI scans for non-red flag headache (a headache without symptoms suggesting a malignant underlying pathology) has more than doubled in the last 10 years in the USA (2000: 6.7% (95% CI 5.4% to 8.2%), 2010: 14% (95% CI 12% to 16%)) (table 1, appendix 6). Conversely, the rate of inappropriate overuse of radiology tests for non-red flag LBP was consistently low, with all (n=18 measures) but two measure showing inappropriate overuse less than 25% of the time (table 1). One of these studies[63] estimated overuse to be about 50% but was conducted in 2001 and thus may reflect improvements over time. The other study is

current but used a small sample size.[64] None of these studies were considered high risk of bias and thus results did not change on further analysis excluding high-risk studies.

6.6.6 Variation of inappropriateness against the same guideline recommendation

Eleven different guideline recommendations were studied more than once. There was significant heterogeneity (I^2 >50%) in nine of these pooled measures. Significant heterogeneity may have occurred for several reasons: (1) vastly different populations (for instance, one study measured the inappropriateness of upper endoscopy in Saudi Arabia[55] using the American Gastroenterological Association recommendations, whereas another study used the same recommendations in the USA)[65]; (2) contrasting healthcare systems,[66,67] (3) relevance and applicability of one country's national guideline to another country;[68] (4) a low number of measures for meta-analysis;[41] and/or (5) significant heterogeneity, reflecting significant variation in inappropriate ordering.

6.7 Discussion

6.7.1 Summary of findings

There is marked variation in the rate of underuse and overuse of diagnostic tests from many primary care settings across the world. This variation suggests improvement can be made in the rate of appropriate diagnostic test ordering.

Primary care use of echocardiography is consistently poor. Echocardiography is inappropriately underused for some clinical situations, for example, confirming a diagnosis of heart failure, and inappropriately overused in others, for example, perioperative assessment. This was consistent across the countries where appropriateness of echocardiogram has been studied. This is of concern given the expertise and resource requirements to perform the test and the increasing availability of direct access ordering for primary care physicians.

For four tests, I found marked variation in the rate of inappropriate use. Underuse of PFTs varied by >50%, whereas overuse of urinary cultures, upper endoscopy and colonoscopy all varied by around 40%.

Radiology tests for both non-red flag LBP and non-red flag headache were frequently *not* overused, but the rate of overuse of imaging for non-red flag headache showed concerning trends, more than doubling from 2000 to 2010.

6.7.3 Strengths and limitations in relation to other studies

Compared with other studies of inappropriate use of healthcare resources, I used data from real clinical encounters. This allowed a more robust assessment of diagnostic test inappropriateness, where other studies used surveys and hypothetical clinical vignettes.[12,69,70] Furthermore, I quantified the appropriateness of all types of diagnostic tests, rather than focusing on a specific test or specific disease (such as only laboratory tests).[22] I present the first systematic review of studies that measured inappropriateness of all diagnostic tests ordered from primary care. Zhi et al[22] quantified the mean rates of overuse and underuse of laboratory tests in secondary care and focused on quantifying an overall rate of overuse and underuse. They estimated that overuse and underuse of laboratory tests was around 21% and 45%, respectively. I chose not to quantify an overall rate of overuse and underuse because I felt combining this data would be inappropriate. It would involve pooling data from different countries and ultimately would not be representative of current practice in any settings.

The use of guidelines as a metric of appropriateness allowed a direct measure of diagnostic test appropriateness. As I have discussed throughout this thesis, assessments of temporal and between-practice variation are useful metrics to highlight potential over and underuse, but they do not consider individual patient data (IPD) or the indication of test use. An assessment of test use against guidelines considers the indication of test use and the patient data; this allows a direct quantification of the amount of test use that is appropriate and inappropriate.

However, the use of guidelines to quantify appropriateness of diagnostic tests could be considered a limitation of this study. Guidelines are often criticised for varying quality[19–21,71] and panel members' conflicts of interests.[72] However, clinical practice guidelines have been shown to improve both care outcomes and processes of care,[18] allow assessment of care on a population level, inform health policy,[73,74] set the standard of care across many healthcare settings[15,16] and provide a medicolegal framework.[17] One major medical insurance company advises that 'doctors must be prepared to explain and justify their decisions and actions, especially if they depart from guidelines produced by a nationally recognised body'.[17] Furthermore, guidelines have been used to measure appropriateness of the use of tests in other published peer-reviewed studies.[22] There will always be times when it is appropriate to depart from guidelines, but marked and consistent variation from guidelines requires investigation. Nevertheless, I investigate the robustness of using guidelines to directly audit test use in Chapter 7.

Furthermore, my study includes only a selection of diagnostic tests and is thus not an all-encompassing reflection of clinical practice. The data I present reflects the use of a specific test in a particular clinical situation in a particular country's healthcare system. Thus, policy makers and those interested in improving the quality of primary care diagnostic test use can use my results as a resource to identify tests in their healthcare setting that require improvement and/or investigation to decipher why such deviation from guidelines exists. The conclusions from this paper, however, are not generalisable to all primary care settings nor all primary care diagnostic tests. Similarly, the data I include represents a snap shot of clinical practice at a certain point in time, often from small centres (e.g. one tertiary centre or from one general practice). Ideally data that more appropriately reflects a nation's or region's appropriate use of a test would lead to potentially more robust conclusions.

It should also be noted that future studies that want to quantify over and underuse on a national level may be limited by the quality of primary care patient records. It is unlikely that current GP patient records will be able to consistently classify the use of a test as inappropriate or appropriate with confidence. When there is clear and succinct indications and contraindications it is likely to be

feasible, for instance for the use of imaging for non-red flag low back pain. However, in situations when there are numerous indications and contraindications it is likely that GPs may write free text to justify the use of lack of use for a test. Thus, on a national-level, it may not be feasible for researchers to look through the free test of millions of patient records. As electronic health records improve, hopefully this will improve.

Lastly, caution must be taken when comparing results that measured inappropriateness using different denominators. The results from studies that measured inappropriateness using patients who had undergone a diagnostic test as a denominator should be interpreted differently to studies that used patients with a diagnosis or symptoms as a denominator (and vice versa).

Further limitations include only searching for studies after 1999, and only searching two medical databases. I chose to only include studies after 1999 for two reasons:

1. The late 90's saw the emergence of evidence-based medicine (EBM) and, in the UK, the established of the National Institute for Clinical Excellence (NICE). I anticipated, a prori, that guidelines would have improved dramatically from 1999.
2. Evidence and guidelines are constantly being updated. Guidelines produced before 1999 are less likely to reflect current best practice.

6.7.2 Implications

6.7.2.1 Implications for policy

Two principle conclusions can be drawn from my results: (1) ordering of echocardiograms from primary care appears to require improvement and (2) markedly varying rates of inappropriate use for PFTs (underuse), colonoscopy (overuse), upper endoscopy (overuse) and urinary cultures (overuse) suggest that ordering can be improved.

6.7.2.2 Implications for clinicians

I have identified that the use of echocardiography from primary care is often not in line with guidelines. It may be of value for primary care clinicians to either a) refamiliarize themselves with the guidelines or b) if they disagree with the guidelines, it would be advantageous for their points to be fed back to guideline developers.

6.7.2.3 Implications for patients

Patients may be surprised to learn that a substantial proportion of primary care test use in incongruent with national or international guidelines. As I have mentioned throughout this chapter, there will certainly be times when it is appropriate to deviate from guidelines, but in these circumstances, it may be pertinent for clinicians and patients to discuss reasons from deviating from guidelines. Shared decision making is particularly important considering the medicolegal advice regarding guideline adherence.

6.7.2.4 Implications for future research

Future research should focus on determining the reasons for deviation from guidelines and an assessment of the quality of guidelines supporting diagnostic test use. Furthermore, investigators wishing to undertake primary studies measuring inappropriate use should focus on developing objective data extraction methods for assessing patient notes and define clearly the interval they (investigators) will consider a test ordered for a particular symptom or disease.

Also, my study only included studies concerning diagnostic tests. Future research can address the amount of overuse and underuse of screening and monitoring tests.

6.7.6 How this chapter fits within my wider thesis

My thesis explores different ways to assess primary care test use. Chapter 3 identifies tests that have been subject to the greatest temporal change in their use, and chapter 4 identifies tests that have been subject to the greatest between-practice variation in their use. In chapter 5 I identify general practices that order a higher proportion of tests that return a normal result. These three chapters present indirect population-level metrics to assess primary care test use. These metrics highlight variation that may

benefit from further assessment; ideally, further assessment should consider the IPD of the patient that underwent a test and also the indications for test use. In this chapter I present one method to achieve this: the assessment of diagnostic test use against a relevant guideline. I have presented a systematic review which captures all the available primary studies which measure the appropriateness of test use against a relevant guideline. In chapter 7 I explore the robustness of guidelines as a criterion in which to measure the appropriateness of test use and in chapter 8 I discuss how my methods can be used in a step-wise approach to identify potentially inappropriate primary care test use.

6.8 Conclusion

There is marked variation in underuse and overuse of appropriate diagnostic test use in primary care across the world. From the available data, echocardiograms are ordered particularly poorly, while the substantial variation in appropriate ordering of PFTs, colonoscopy, upper endoscopy and urinary cultures suggests a need for improvement.

References

1 O'Sullivan JW, Albasri A, Nicholson B, *et al.* Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open* 2018; **8**: e018557.

2 Epner PL, Gans JE, Graber ML. When diagnostic testing leads to harm: a new outcomes-based approach for laboratory medicine. *BMJ Qual Saf* 2013; **22 Suppl 2**: ii6-ii10.

3 Gandhi TK, Kachalia A, Thomas EJ, Puopolo AL, Yoon C. Missed and Delayed Diagnoses in the Ambulatory Setting. *Ann Intern Med* 2006; **145**: 488–96.

4 Katzberg RW, Lamba R. Contrast-induced nephropathy after intravenous administration: fact or fiction? *Radiol Clin North Am* 2009; **47**: 789–800, v.

5 Lumbreras B, Donat L, Hernández-Aguado I. Incidental findings in imaging diagnostic tests: a systematic review. *Br J Radiol* 2010; **83**: 276–89.

6 Brodersen J, Schwartz LM, Heneghan C, O'Sullivan JW, Aronson J, Woloshin S. Overdiagnosis: what it is and what it isn't. *BMJ EBM* 2018; **23**: 1–3.

7 Moynihan R, Doust J, Henry D. Preventing overdiagnosis: how to stop harming the healthy. *Bmj* 2012; **344**: e3502–e3502.

8 Berwick D, Hackbarth AD. Eliminating Waste in US Health Care. *JAMA* 2012; **307**: 1513.

9 Cecchini M, Lee S, For the Organisation for Economic Co-operation and Development. Tackling Wasteful Spending on Healthcare. 2017 http://networks.sustainablehealthcare.org.uk/sites/default/files/resources/Tackling Wasteful Spending on Health.pdf#page=117.

10 Department of Health. NHS 2010–2015: from good to great. preventative, people-centred, productive. London, 2009.

11 Esmail A, Neale G, Elstein M, Firth-Cozens J, Davy C VC. Case Studies in Litigation: Claims reviews in four specialties. Manchester, 2004.

12 Sirovich BE, Woloshin S, Schwartz LM. Too Little? Too Much? Primary care physicians' views on US health care: a brief report. *Arch Intern Med* 2011; **171**: 1582–5.

13 Hobbs FDR, Bankhead C, Mukhtar T, *et al.* Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14. *Lancet* 2016; **387**: 2323–30.

14 Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? *PLoS Med* 2010; **7**. DOI:10.1371/journal.pmed.1000326.

15 Garber AM. Evidence-based guidelines as a foundation for performance incentives. *Health Aff (Millwood)* 2005; **24**: 174–9.

16 Ransohoff DF, Pignone M, Sox HC, *et al.* How to Decide Whether a Clinical Practice Guideline Is Trustworthy. *JAMA* 2013; **309**: 139.

17 Fryar C. Doctors can depart from guidelines in patients' best interests. *BMJ* 2015; **350**.

18 Grimshaw JM, Russell IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. *Lancet (London, England)* 1993; **342**: 1317–22.

19 Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature.

JAMA 1999; **281**: 1900–5.

20 Grilli R, Magrini N, Penna A, *et al.* Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet (London, England)* 2000; **355**: 103–6.

21 Lenzer J. Why we can't trust clinical guidelines. *BMJ* 2013; **346**.

22 Zhi M, Ding EL, Theisen-Toupal J, *et al.* The Landscape of Inappropriate Laboratory Testing: A 15-Year Meta-Analysis. *PLoS One* 2013; **8**: e78962.

23 McGlynn EA, Asch SM, Adams J, *et al.* Quality of health care delivered to adults in the United States. *N Engl J Med* 2003; **349**: 1866-8-8.

24 Sheldon T a, Cullum N, Dawson D, *et al.* What's the evidence that NICE guidance has been implemented? Results from a national evaluation using time series analysis, audit of patients' notes, and interviews. *BMJ* 2004; **329**: 999.

25 Alderwick H, Robertson R, Appleby J, Dunn P, Maguire D. Better value in the NHS The role of changes in clinical practice. 2015.

26 Fisher ES, Bynum JP, Skinner JS. Slowing the growth of health care costs--lessons from regional variation. *N Engl J Med* 2009; **360**: 849–52.

27 Appleby J, Thompson J, Jabbal J. Quarterly Monitoring Report: How is the NHS performing? *King's Fund* 2016; : 1–42.

28 Foot C, Naylor C, Imison C. The quality of GP diagnosis and referral. 2010 http://amapro.isabelhealthcare.com/pdf/Kings_Fund_Diagnosis_and_Referral_2010.pdf.

29 Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009; **339**: b2535.

30 Wald NJ. Guidance on terminology. *J Med Screen* 2008; **15**: 50–50.

31 Raffle A, Gray J. Screening: Evidence and Practice. Oxford University Press, 2007.

32 Glasziou P, Irwig L, Aronson J. Evidence-based medical monitoring: from principles to practice. Oxford (UK): Blackwell Publishing, BMJ books, 2008.

33 Hoy D, Brooks P, Woolf A, *et al.* Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement. *J Clin Epidemiol* 2012; **65**: 934–9.

34 Belletti D, Liu J, Zacker C, Wogen J. Results of the CAPPS: COPD--assessment of practice in primary care study. *Curr Med Res Opin* 2013; **29**: 957–66.

35 Bertella E, Zadra A, Vitacca M, General Practitioners Distretto 04 ASL 02 Brescia, Italy. COPD management in primary care: is an educational plan for GPs useful? *Multidiscip Respir Med* 2013; **8**: 24.

36 Chavez PC, Shokar NK. Diagnosis and management of chronic obstructive pulmonary disease (COPD) in a primary care clinic. *COPD* 2009; **6**: 446–51.

37 Lange P, Rasmussen FV, Borgeskov H, *et al.* The quality of COPD care in general practice in Denmark: the KVASIMODO study. *Prim Care Respir J* 2007; **16**: 174–81.

38 Ulrik CS, Sørensen TB, Højmark TB, Olsen KR, Vedsted P. Adherence to COPD guidelines in general practice: impact of an educational programme delivered on location in Danish general practices. *Prim Care Respir J* 2013; **22**: 23–8.

39 Barendregt JJ, Doi SA, Lee YY, Norman RE, Vos T. Meta-analysis of prevalence. *J Epidemiol Community Heal* 2013; : 974–8.

40 Doi SAR, Barendregt JJ, Khan S, Thalib L, Williams GM. Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model. *Contemp Clin Trials* 2015; **45**: 130–8.

41 Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; **327**: 557–60.

42 Mafi JN, McCarthy EP, Davis RB, Landon BE. Worsening Trends in the Management and Treatment of Back Pain. *JAMA Intern Med* 2013; **173**: 1573.

43 Mafi JN, Edwards ST, Pedersen NP, Davis RB, McCarthy EP, Landon BE. Trends in the Ambulatory Management of Headache: Analysis of NAMCS and NHAMCS Data 1999-2010. *J Gen Intern Med* 2015; **30**: 548–55.

44 Williams CM, Maher CG, Hancock MJ, *et al.* Low Back Pain and Best Practice Care. *Arch Intern Med* 2010; **170**: 271.

45 Cai JX, Campbell EJ, Richter JM. Concordance of Outpatient Esophagogastroduodenoscopy of the Upper Gastrointestinal Tract With Evidence-Based Guidelines. *JAMA Intern Med* 2015; **175**: 1563–4.

46 Gurzun M-M, Ionescu A. Appropriateness of use criteria for transthoracic echocardiography: are they relevant outside the USA? *Eur Hear J - Cardiovasc Imaging* 2014; **15**: 450–5.

47 van Gurp N, Boonman-De winter LJM, Meijer Timmerman Thijssen DW, Stoffers HEJH.

Benefits of an open access echocardiography service: A Dutch prospective cohort study. *Netherlands Hear J* 2013; **21**: 399–405.

48 Johnson JD, O'Mara HM, Durtschi HF, Kopjar B. Do Urine Cultures for Urinary Tract Infections Decrease Follow-up Visits? *J Am Board Fam Med* 2011; **24**: 647–55.

49 Grover ML, Bracamonte JD, Kanodia AK, *et al.* Assessing Adherence to Evidence-Based Guidelines for the Diagnosis and Management of Uncomplicated Urinary Tract Infection. *Mayo Clin Proc* 2007; **82**: 181–5.

50 Llor C, Rabanaque G, Lopez A, Cots JM. The adherence of GPs to guidelines for the diagnosis and treatment of lower urinary tract infections in women is poor. *Fam Pract* 2011; **28**: 294–9.

51 Lindbäck H, Lindbäck J, Melhus Å. Inadequate adherence to Swedish guidelines for uncomplicated lower urinary tract infections among adults in general practice. *Apmis* 2017; **125**: 816–21.

52 Leon P, Catherine K, Mark N, John D. Gastro-oseophageal reflux disease. The impact of guidelines on GP management. 2008.

53 Hughes-Anderson W, Rankin SL, House J, Aitken J, Heath D, House AK. Open access endoscopy in rural and remote Western Australia: does it work? *ANZ J Surg* 2002; **72**: 699–703.

54 Aljebreen AM, Alswat K, Almadi MA. Appropriateness and diagnostic yield of upper gastrointestinal endoscopy in an open-access endoscopy system. *Saudi J Gastroenterol* 2013; **19**: 219–22.

55 Azzam NA, Almadi MA, Alamar HH, *et al.* Performance of American Society for Gastrointestinal Endoscopy guidelines for dyspepsia in Saudi population: Prospective observational study. *World J Gastroenterol* 2015; **21**: 637–43.

56 Elwyn G, Owen D, Roberts L, *et al.* Influencing referral practice using feedback of adherence to NICE guidelines: a quality improvement report for dyspepsia. *Qual Saf Health Care* 2007; **16**: 67–70.

57 Cardin F, Zorzi M, Bovo E, *et al.* Effect of Implementation of a Dyspepsia and Helicobacter pylori Eradication Guideline in Primary Care. *Digestion* 2005; **72**: 1–7.

58 Hassan C, Bersani G, Buri L, *et al.* Appropriateness of upper-GI endoscopy: an Italian survey on behalf of the Italian Society of Digestive Endoscopy. *Gastrointest Endosc* 2007; **65**: 767–74.

59 Cardin F, Zorzi M, Terranova O. Implementation of a guideline versus use of individual prognostic factors to prioritize waiting lists for upper gastrointestinal endoscopy. *Eur J Gastroenterol Hepatol* 2007; **19**: 549–53.

60 Fiorenza JP, Tinianow AM, Chan WW. The Initial Management and Endoscopic Outcomes of Dyspepsia in a Low-Risk Patient Population. *Dig Dis Sci* 2016; **61**: 2942–8.

61 Chan Y-M, Goh K-L. Appropriateness and diagnostic yield of EGD: a prospective study in a large Asian hospital. *Gastrointest Endosc* 2004; **59**: 517–24.

62 CHAN T, GOH K. Appropriateness of colonoscopy using the ASGE guidelines: experience in a large Asian hospital. *Chin J Dig Dis* 2006; **7**: 24–32.

63 Eccles M, Steen N, Grimshaw J, *et al.* Effect of audit and feedback, and reminder messages on primary-care radiology referrals: a randomised trial. *Lancet* 2001; **357**: 1406–9.

64 Tahvonen P, Oikarinen H, Niinimäki J, Liukkonen E, Mattila S, Tervonen O. Justification and active guideline implementation for spine radiography referrals in primary care. *Acta radiol* 2016; **58**: 586–92.

65 Majumdar SR, Soumerai SB, Farraye FA, *et al.* Chronic acid-related disorders are common and underinvestigated. *Am J Gastroenterol* 2003; **98**: 2409–14.

66 Basu S, Andrews J, Kishore S, Panjabi R, Stuckler D. Comparative performance of private and public healthcare systems in low- and middle-income countries: A systematic review. *PLoS Med* 2012; **9**: 19.

67 Ridic G, Gleason S, Ridic O. Comparisons of Health Care Systems in the United States , Germany and Canada. *Mat Soc Med* 2012; **24**: 112–20.

68 Gagliardi AR, Brouwers MC. Do guidelines offer implementation advice to target users? A systematic review of guideline applicability. *BMJ Open* 2015; **5**: e007047–e007047.

69 Kachalia A, Berg A, Fagerlin A, *et al.* Overuse of testing in preoperative evaluation and syncope: a survey of hospitalists. *Ann Intern Med* 2015; **162**: 100–8.

70 Swennen MHJ, Rutten FH, Kalkman CJ, van der Graaf Y, Sachs APE, van der Heijden GJMG. Do general practitioners follow treatment recommendations from guidelines in their decisions on heart failure management? A cross-sectional study. *BMJ Open* 2013; **3**: e002982.

71 Burgers JS, Fervers B, Haugh M, *et al.* International Assessment of the Quality of Clinical

Practice Guidelines in Oncology Using the Appraisal of Guidelines and Research and Evaluation Instrument. *J Clin Oncol* 2004; **22**: 2000–7.

72 Gale EAM. Conflicts of interest in guideline panel members. *BMJ* 2011; **343**.

73 Institute of Medicine. Clinical Practice Guidelines: Directions for a New Program. Washington, 1990 DOI:10.1097/SPV.0b013e31828a2951.

74 Browman GP, Snider A, Ellis P. Negotiating for change. The healthcare manager as catalyst for evidence-based practice: changing the healthcare environment and sharing experience. *Healthc Pap* 2003; **3**: 10–22.

Chapter Seven An assessment of the robustness of guidelines for auditing test use

7.1 Publication statement

A paper based on this chapter has been published in the Journal of Clinical Epidemiology.

7.2 Introduction

In Chapter 6 (*Overtesting and undertesting in primary care: a systematic review*) I explored the extent of primary care physician's non-adherence to diagnostic test guideline recommendations. I considered non-adherence to have occurred when tests were ordered against a guideline recommendation (overuse), and when tests were not ordered when a guideline recommended to order (underuse). The resulting measures of diagnostic test non-adherence were the percentages of tests that were either over or underused, measured against their respective guidelines. These results are presented in Chapter 6.

Currently, audit against a guideline or an appropriateness criterion is one of the only way to directly quantify overuse and underuse of tests.[1] However, using guidelines to directly audit test use assumes that guidelines are produced appropriately and that their recommendations are based on high-quality evidence.

These assumptions are not always correct. Guidelines have been criticized for varying quality, particularly, biases in their development[2,3] and lack or poor quality evidence supporting their recommendations.[4] These limitations jeopardise the use of guidelines for directly auditing the appropriateness of test use.

In this chapter I test the robustness of guidelines for auditing test use. To do this, I assessed the quality and reporting of guidelines *and* the quality of the evidence supporting guideline recommendations.

Terminology

- *Measures of non-adherence:* the percentage of tests that were either overused or underused, measured against their respective guideline. These measures are taken from primary studies in Chapter 6.
- *Quality and reporting of guidelines:* Methods used to construct a guideline and the reporting standards of a guidelines, measured using the AGREE II tool.
- *Evidence quality*: Assessment of the evidence underpinning guideline recommendations using GRADE

7.3 Aim

To test the robustness of guidelines in auditing the use of diagnostic tests.

7.4 Objectives

1. To determine the quality and reporting of guidelines used to assess diagnostic test ordering in *Chapter 6*
2. To determine the quality of evidence underpinning the guidelines used to assess diagnostic test ordering in *Chapter 6*
3. To examine the relationship between the quality and reporting of guidelines (AGREE II score) and guideline non-adherence
4. To examine the relationship between the quality of evidence underpinning diagnostic test guidelines (GRADE score) and guideline non-adherence

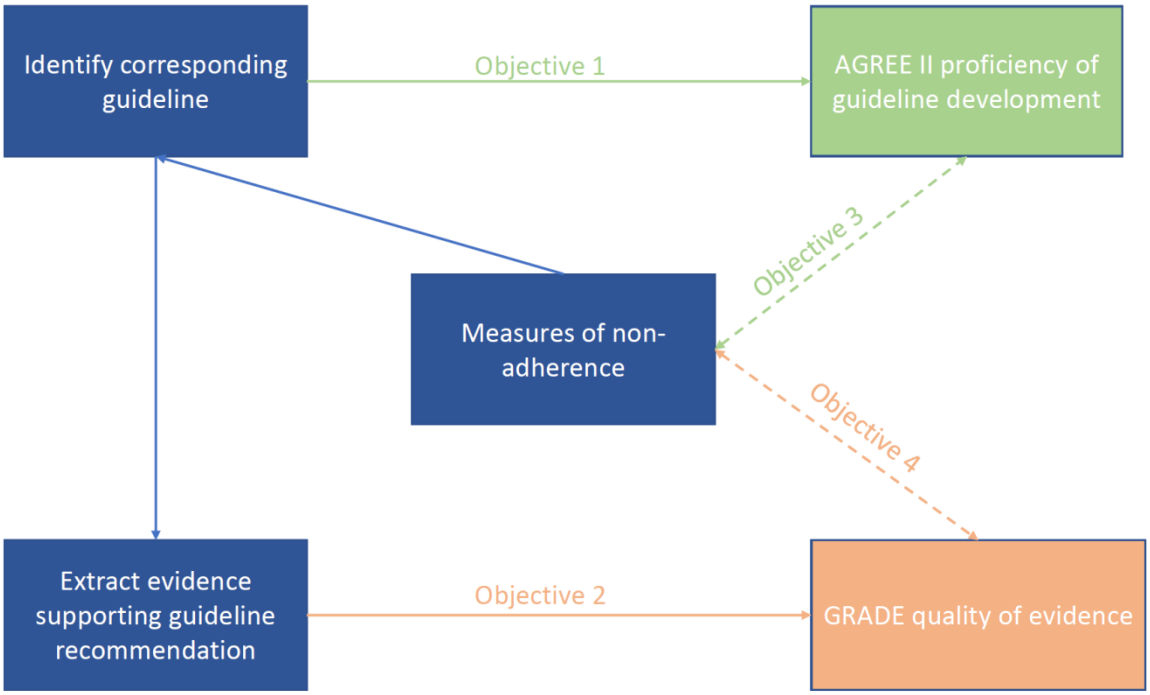


Figure 7.1 Methods flow diagram

7.5 Methods

7.5.1 Study design

Five steps defined the conduct of this study (Figure 7.1):

1. *Measures of non-adherence:* The measures of guideline non-adherence were extracted from *Chapter 6.*
2. *Assessment of the quality and reporting of guidelines:* For each of the measures of non-adherence, I identified the respective guidelines against which adherence was measured. I then used the AGREE II tool to determine the quality and reporting of these guidelines.

3. *Assessment of evidence quality:* For each identified guideline (from Step 2), I identified the evidence supporting the guideline recommendations. I then assessed the quality of the evidence using GRADE.
4. *Association between guideline non-adherence and quality and reporting of a guideline (AGREE II):* I constructed linear models to examine the relationship between the quality and reporting of guidelines (Step 2) and measures of guideline non-adherence (Step 1).
5. *Association between guideline non-adherence and evidence quality (GRADE):* I constructed linear models to examine the relationship between evidence quality (Step 3) and measures of guideline non-adherence (Step 1).

7.5.2 Guideline non-adherence

Chapter 6 is a systematic review that determined the non-adherence of diagnostic test ordering against their respective guidelines. The full methods are reported in Chapter 6 and its corresponding publication.[5] Briefly, I searched three databases (MEDLINE, EMBASE, and the Cochrane Library) and included primary observational studies that measured the non-adherence of diagnostic test ordering against national or international guidelines. I extracted each primary study's measure of non-adherence. These measures of non-adherence, along with a description of the type of diagnostic test, the respective guideline, and the relevant recommendation, are listed in Appendix 6 (and Tables 1 and 2 of its corresponding publication).[5]

7.5.3 Identify guidelines

Chapter 6 describes 103 measures of non-adherent diagnostic test ordering. Each of these measures are from individual studies, which also report the respective guideline used to measure adherence against. I identified each guideline and extracted the following: the guideline recommendation, the full name and reference of the guideline, the professional society that published the guideline (e.g. NICE, American Heart Association), and the year of publication. Guidelines not written in English were excluded.

The full name and reference of each guideline was then used to locate the full guideline in Google, Google Scholar, the Professional Society website (e.g. European Association of Urology), or PubMed. If this search failed, I searched the Professional Society websites; if still unsuccessful, I emailed the Professional Society directly.

7.5.4 Assessment of the quality and reporting of guidelines using the AGREE II tool

From the primary studies included in *Chapter 6*, I identified the guidelines that were used to measure non-adherence against.

Two reviewers (JOS, AA) independently assessed all included guidelines using the AGREE II tool. The AGREE II is a tool that 'assesses the methodological rigour and transparency in which a guideline is developed' [6]. The tool itself consists of 23 questions organised into six domains (Scope and Purpose, Stakeholder involvement, Rigour of Development, Clarity of Presentation, Applicability and Editorial Independence). Each domain is assigned a score of 0% to 100%; where 100% indicates that the guideline scored perfectly for that domain. Aside from the 23 questions, the AGREE II tool also prompts users to generate an overall score for the guideline (Appendix 7.1). The 'overall assessment' is a quality score (1-7) for the entire guideline, where 7 represents 'Highest possible quality' and 1 'Lowest possible quality'.

Two reviewers (JOS, AA) independently assessed all included guidelines using the AGREE II tool, both reviewers generated a score for each of its six domains and an overall score for the guideline. I then used the formula in the AGREE II guidance to generate an aggregate score for each domain (Appendix 7.2). The score is expressed as a percentage of the maximum possible score. The exception to this is the scoring for the 'Overall Assessment', where an average of the two reviewer's scores was used. For the overall assessment score, I performed a weighted kappa test to compare agreement between reviewers.

I completed online training on the AGREE II tool (http://www.agreetrust.org/resource-centre/agree-ii-training-tools/) and attended face-to-face training in the Practice of Evidence-Based Health Care module in the Masters course in Evidence-Based Health-care.

7.5.5 Assessment of evidence quality (GRADE)

From the identified guidelines, I extracted the reference(s) quoted to support specific diagnostic test recommendations. For instance, one of the primary studies in Chapter 6 was Mafi 2013,[7] which measured overuse of imaging tests for patients with non-red flag low back pain. This study measured non-adherence against the guideline recommendation: '*no imaging for non-red flag low back pain'* from the guideline: '*Diagnosis and Treatment of Low Back Pain: A Joint Clinical Practice Guideline from the American College of Physicians and the American Pain Society'*).[8] This American College of Physicians and American Pain Society guideline used the following references to support this recommendation: *Kerry et al Radiography for low back pain: a randomised controlled trial and observational study in primary care. The British Journal of General Practice, 52: 469–474* [9] and *Kendrick et al Radiography of the lumbar spine in primary care patients with low back pain: randomised controlled trial. BMJ, 322: 400–405)* [10]. These studies were extracted and then assessed, as described below.

I searched PubMed and the Bodleian Library of the University of Oxford for the full texts of studies supporting guideline recommendations. If the recommendation was not referenced, the professional societies/organisation or contact author of the guideline was directly contacted via email. In the absence of clarification from the guideline organisation, I assumed recommendations were expert opinion. Some recommendations explicitly stated they were based on 'expert opinion', these were also classed as expert opinion.

Two reviewers assessed the quality of evidence using GRADE.[11] We produced one GRADE assessment for each measure of non-adherence. I made an independent assessment, which was then verified by the second author (AA). The GRADE system is used to 'grade evidence and recommendations'.[11] It consists of five categories, including Risk of Bias ('Study Limitations'),[12]

Inconsistency of results,[13] Indirectness of evidence,[14] Imprecision of evidence[15] and other biases (publication, effect size, confounding, dose-response gradient). Because of uncertainty in assessing publication bias in diagnostic accuracy studies,[16] I did not assess other biases. The assessments within each category then contribute to the overall GRADE score; which is one of high, moderate, low, or very low quality. As recommended by GRADE, if one of the categories was deemed to have serious limitations, the evidence was downgraded by one, e.g. from high to moderate. As is recommended by GRADE,[11] expert opinion was considered 'very low quality' evidence.

GRADE is an outcome-specific assessment, rather than an assessment of individual studies, as is the case with the Cochrane Risk of Bias tool.[17] Thus, GRADE assesses a collection of studies supporting one outcome and does not give a quality assessment of one study's multiple outcomes.

The identified studies underpinning guideline recommendations were of varying designs. When the underpinning studies were diagnostic accuracy studies, or systematic reviews of diagnostic accuracy studies, we (myself and AA) used GRADE for diagnostic tests.[18] Otherwise, we used GRADE for interventions.[11] The two GRADE systems follow the same structure (risk of bias, indirectness, inconsistency, imprecision, and other biases), but are subtly different in their assessments (Table 7.1).

Table 7.1 Differences between GRADE for interventions and GRADE for Diagnostic accuracy[16,18–20]

	Intervention research	**Diagnostic accuracy research**
Basics	One outcome is considered	Two outcomes are considered: sensitivity and specificity
GRADE Categories		
Risk of bias	Cochrane risk of bias tool or Cochrane risk of bias in cohort studies	QUADAS-2
Indirectness	Compare evidence PICO to guideline PICO	Applicability section of QUADAS-2 and an assessment of the directness of sensitivity and specificity to patient outcomes
Inconsistency	I^2 or χ^2 tests for heterogeneity	Visual inspection with statistical tests of heterogeneity interpreted with caution.
Imprecision	95% Confidence intervals	95% Confidence intervals

Risk of bias is one of the five categories in GRADE. We (myself and AA) used different risk of bias tools to assess primary studies with different study designs, as recommended by GRADE.[12] For systematic reviews we used the AMSTAR checklist,[21] and for randomised controlled trials (RCTs) we used the Cochrane risk of bias tool.[17] We used the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) 2 tool[22] and the Cochrane risk of bias in cohort studies tool[23] for assessing diagnostic accuracy studies and cohort studies respectively.

In some instances, a guideline included multiple recommendations pertaining to one specific diagnostic test. For example, The American College of Physicians guideline for upper gastrointestinal endoscopy for Gastroesophageal Reflux Disease[24] recommends ordering a upper endoscopy for any of the following: "1. Nondysplastic Barrett's esophagus without an EGD for surveillance in the past 3 years, 2. Acute symptoms (<5 years) that persisted despite 4 to 8 weeks of twice-daily proton-pump-inhibitor therapy, 3. Chronic symptoms (>5 years) in men who were older than 50 years, 4. Alarm symptoms (dysphagia, bleeding, anemia, weight loss, and recurrent vomiting), and 5. Severe erosive esophagitis despite two months of proton-pump inhibitor therapy." Each of these recommendations was supported by unique evidence (studies) and thus each recommendation had a unique GRADE assessment. Such GRADE assessments could potentially be different from each other, for example the evidence supporting one recommendation (e.g. "Nondysplastic Barrett's esophagus without an EGD for surveillance in the past 3 years") may have a GRADE assessment of High, compared to another recommendation from the same guideline (e.g. "Chronic symptoms (>5 years) in men who were older than 50 years") may have a GRADE assessment of Very low. This was problematic, because I had only one measure of non-adherent diagnostic test ordering for each guideline recommendation. Thus, to represent the data appropriately I used one GRADE assessment per guideline recommendation. Had I used the same measure of non-adherence for each different GRADE assessment, I would have inappropriately duplicated the data. I therefore specified a priori that I would do the following:

1. If GRADE assessments were consistent across all recommendations (e.g. 'moderate quality'), I would use this GRADE assessment.

2. If the GRADE assessments were different across recommendations, I would select the most frequent assessment (e.g. if there were five outcomes and four were of 'low quality', I would choose 'low quality').
3. If the frequency of GRADE assessments were the same, I would, as recommended by GRADE, select the patient-oriented outcome, rather than diagnostic accuracy or another surrogate.[18]

Similarly, the GRADE system for diagnostic tests produces a GRADE score for both sensitivity and specificity. This was similarly problematic. If the GRADE score differed between sensitivity and specificity, I chose the more clinically relevant measure. I determined clinically relevance by consulting the guideline to determine if the test was recommended to rule in or rule out disease. If it was used to rule in, I chose specificity; if it was used to rule out, I chose sensitivity.

7.5.6 Linear Regression models

I constructed the following linear models:

Model A: Percentage guideline non-adherence (0-100%) ~ f(quality and reporting of a guideline (Overall AGREE II Score, 1-7))

Model B: Percentage guideline non-adherence (0-100%) ~ f(evidence quality (GRADE: Very low, low, moderate or high))

7.5.6.1 Model A: Association between quality and reporting of a guideline (AGREE II) and guideline non-adherence

I constructed six further models to assess the relationship between guideline non-adherence and the domains of AGREE II.

Model A.1: Percentage guideline non-adherence (0-100%) ~ f(Domain 1 AGREE II score (0-100%)

Model A.2: Percentage guideline non-adherence (0-100%) ~ f(Domain 2 AGREE II score (0-100%)

Model A.3: Percentage guideline non-adherence (0-100%) ~ f(Domain 3 AGREE II score (0-100%)

Model A.4: Percentage guideline non-adherence (0-100%) ~ f(Domain 4 AGREE II score (0-100%)

Model A.5: Percentage guideline non-adherence (0-100%) ~ f(Domain 5 AGREE II score (0-100%)

Model A.6: Percentage guideline non-adherence (0-100%) ~ f(Domain 6 AGREE II score (0-100%)

7.5.6.2 Model B: Association between Evidence quality (GRADE) and Guideline non-adherence

For model B, GRADE categories were converted into numeric categories (Very low = 1, Low = 2, Moderate = 3, and High = 4). Model B was a linear regression model with a categorical predictor.

7.5.6.3 Sensitivity analyses

I performed a sensitivity analysis based on the quality of the primary studies from Chapter 6. As previously described, these primary studies provided the measures of non-adherence. In Chapter 6, these studies were rated as low, moderate or high risk of bias, in accordance to the risk of bias tool outlined by Hoy et al.[25] The following models were constructed:

1. For primary studies classified as *low* risk of bias:
    a. Percentage guideline non-adherence (0-100%) ~ f(guideline quality and reporting (Overall AGREE II Score, 1-7))
    b. As I did in the main analysis, I created six models for each of the six domains of the AGREE II tool, i.e.
        i. Percentage guideline non-adherence (0-100%) ~ f(Domain 1 II Score (1-7))
        ii. Percentage guideline non-adherence (0-100%) ~ f(Domain 2 II Score (1-7))
        iii. Percentage guideline non-adherence (0-100%) ~ f(Domain 3 II Score (1-7))
        iv. Percentage guideline non-adherence (0-100%) ~ f(Domain 4 II Score (1-7))
        v. Percentage guideline non-adherence (0-100%) ~ f(Domain 5 II Score (1-7))
        vi. Percentage guideline non-adherence (0-100%) ~ f(Domain 6 II Score (1-7))
    c. Percentage guideline non-adherence (0-100%) ~ f(evidence quality (GRADE: Very low, low, moderate or high))

2. For primary studies classified as *moderate* risk of bias:
    a. Percentage guideline non-adherence (0-100%) ~ f(guideline quality and reporting (Overall AGREE II Score, 1-7))
    b. As we did in the main analysis, we created six models for each of the six domains of the AGREE II tool, i.e.
        i. Percentage guideline non-adherence (0-100%) ~ f(Domain 1 II Score (1-7))
        ii. Percentage guideline non-adherence (0-100%) ~ f(Domain 2 II Score (1-7))
        iii. Percentage guideline non-adherence (0-100%) ~ f(Domain 3 II Score (1-7))
        iv. Percentage guideline non-adherence (0-100%) ~ f(Domain 4 II Score (1-7))
        v. Percentage guideline non-adherence (0-100%) ~ f(Domain 5 II Score (1-7))
        vi. Percentage guideline non-adherence (0-100%) ~ f(Domain 6 II Score (1-7))
    c. Percentage guideline non-adherence (0-100%) ~ f(evidence quality (GRADE: Very low, low, moderate or high))
3. For primary studies classified as *high* risk of bias:
    a. Percentage guideline non-adherence (0-100%) ~ f(guideline quality and reporting (Overall AGREE II Score, 1-7))
    b. As we did in the main analysis, we created six models for each of the six domains of the AGREE II tool, i.e.
        i. Percentage guideline non-adherence (0-100%) ~ f(Domain 1 II Score (1-7))
        ii. Percentage guideline non-adherence (0-100%) ~ f(Domain 2 II Score (1-7))
        iii. Percentage guideline non-adherence (0-100%) ~ f(Domain 3 II Score (1-7))
        iv. Percentage guideline non-adherence (0-100%) ~ f(Domain 4 II Score (1-7))
        v. Percentage guideline non-adherence (0-100%) ~ f(Domain 5 II Score (1-7))
        vi. Percentage guideline non-adherence (0-100%) ~ f(Domain 6 II Score (1-7))
    c. Percentage guideline non-adherence (0-100%) ~ f(evidence quality (GRADE: Very low, low, moderate or high))

7.5.6.4 Subgroup analysis

I performed an additional 12 subgroup analyses where the measures of guideline non-adherence ("percentage guideline non-adherence") were stratified into overuse (the ordering of a test when it is not recommended) or underuse (the failure to order a test when it is recommended). I constructed the following models:

Underuse

1. Percentage guideline non-adherence (0-100%) ~ f(AGREE II Score (1-7))
2. As I did in the main analysis, I created six models for each of the six domains of the AGREE II tool, i.e.
    a. Percentage guideline non-adherence (0-100%) ~ f(Domain 1 II Score (1-7))
    b. Percentage guideline non-adherence (0-100%) ~ f(Domain 2 II Score (1-7))
    c. Percentage guideline non-adherence (0-100%) ~ f(Domain 3 II Score (1-7))
    d. Percentage guideline non-adherence (0-100%) ~ f(Domain 4 II Score (1-7))
    e. Percentage guideline non-adherence (0-100%) ~ f(Domain 5 II Score (1-7))
    f. Percentage guideline non-adherence (0-100%) ~ f(Domain 6 II Score (1-7))
3. Percentage guideline non-adherence (0-100%) ~ f(GRADE Score (Very low, low, moderate or high))

Overuse

1. Percentage guideline non-adherence (0-100%) ~ f(AGREE II Score (1-7))
2. As I did in the main analysis, I created six models for each of the six domains of the AGREE II tool, i.e.
    a. Percentage guideline non-adherence (0-100%) ~ f(Domain 1 II Score (1-7))
    b. Percentage guideline non-adherence (0-100%) ~ f(Domain 2 II Score (1-7))
    c. Percentage guideline non-adherence (0-100%) ~ f(Domain 3 II Score (1-7))
    d. Percentage guideline non-adherence (0-100%) ~ f(Domain 4 II Score (1-7))
    e. Percentage guideline non-adherence (0-100%) ~ f(Domain 5 II Score (1-7))
    f. Percentage guideline non-adherence (0-100%) ~ f(Domain 6 II Score (1-7))

3. Percentage guideline non-adherence (0-100%) ~ f(GRADE Score (Very low, low, moderate or high))

For all models, the distributions of the dependent and independent variables was determined separately. Model assumptions were checked and in the case of violations appropriate transformations were applied.

For all models, coefficients have been determined and are presented with corresponding 95% confidence intervals and their P values. All analyses were performed in R version 3.4.1.

7.6 Results

7.6.1 Identifying and locating guidelines

I extracted 103 measures of non-adherent testing from Chapter 6. Of these 103 measures of non-adherent testing, I included 62. Measures of non-adherent testing were excluded for several reasons (Figure 7.2).

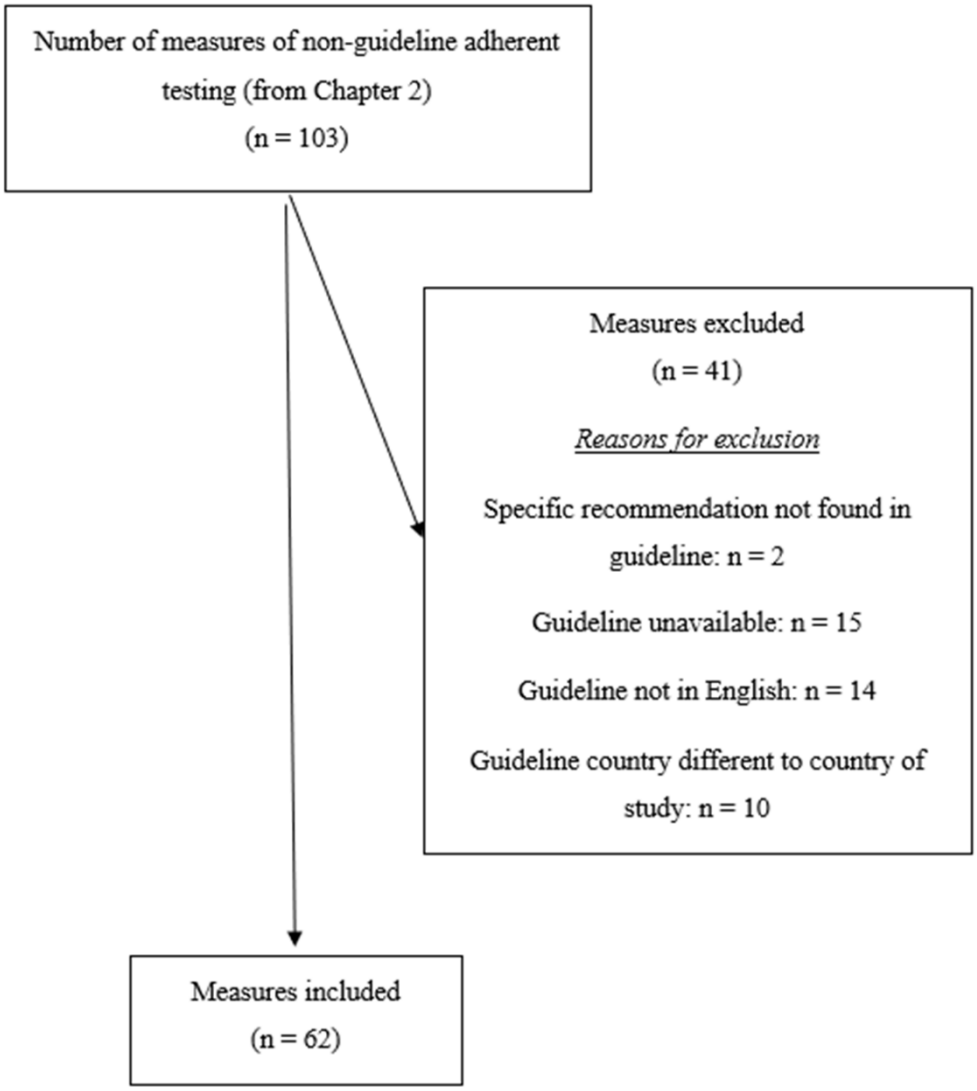


Figure 7.2 Inclusion and exclusion of measures of non-adherent testing

7.6.2 Assessment of the quality and reporting of guidelines using AGREE II

Appendix Table 7.1 lists the AGREE II scores for each of the guidelines used for the 62 measures of non-adherence. The median score for the overall assessment was 4.5 (IQR: 3.5 to 5). The weighted kappa between the two reviewers for the 'Overall Assessment' score was 0.77.

The first AGREE II domain (*Scope and Purpose of the guideline)* was the domain with the highest score across guidelines (median score: 92%), whereas Domain 5 (*Applicability of the guideline)* had the lowest (median score: 33%). The median percentage scores for each domain were: domain 1

'Scope and purpose' 92% (IQR: 81% to 97%), domain 2: 'Stakeholder involvement' 44% (IQR: 37% to 66%), domain 3: 'Rigour of development': 60% (IQR: 29% to 84%), domain 4: 'Clarity of presentation' 78% (IQR: 75% to 92%), domain 5: 'Applicability' 33% (IQR: 13% to 44%) and domain 6: 'Editorial Independence' 50% (IQR: 13% to 67%).

7.6.3 Assessment of evidence quality (GRADE)

Most evidence (n=45, 73%) were graded as very low, with seven low (11%), and ten high (16%) (none were graded as moderate). Of the 45 measures graded as very low-quality evidence, 37 were expert opinion. The GRADE assessment for the 25 guideline recommendations based on evidence (rather than expert opinion) are displayed in Appendix Table 7.2. For nine of these 25 GRADE assessments we used GRADE for diagnostic tests and, for the other 16 we used GRADE for interventions.

7.6.4 Association between guideline adherence and the quality and reporting of guidelines (AGREE II)

7.6.4.1 Model A: Percentage guideline non-adherence (0-100%) ~ f(quality and reporting of guidelines (Overall AGREE II Score, 1-7))

The linear regression model showed no significant association between the quality and reporting of guidelines (AGREE II score) and the percentage of non-adherent testing. As guideline quality and reporting improved, there was a non-significant fall in non-adherent testing (coefficient: -4.2% (95%CI: -9.0% to 0.7%, P = 0.09). Figure 7.3 shows the median percentage of non- adherent testing for each AGREEII category.

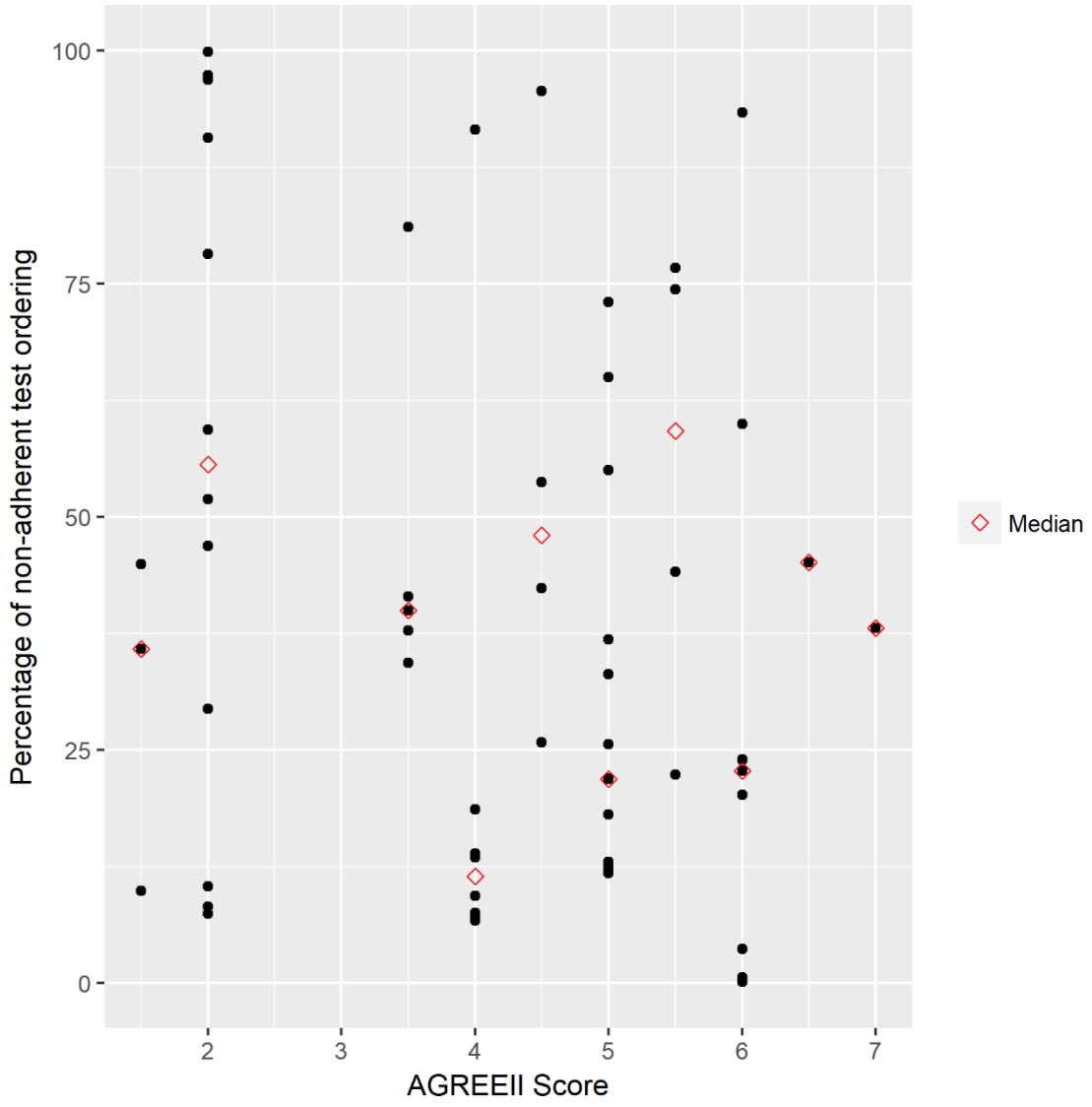


Figure 7.3 Scatter plot and median guideline quality and reporting (Overall AGREE II Score) and % guideline non-adherence

7.6.4.2 Model A.1 to A.6: Association between AGREE II domains and guideline non-adherence

I explored the associations between the scores in the six domains of the AGREE II tool and the percentages of non-adherent testing using linear regression models (models A.1 to A.6). High scores in three domains of the AGREE II tool were significantly associated with reductions in non-adherent testing (Domain 2: 'Stakeholder involvement', Domain 3: 'Rigour of development' and Domain 5: 'Applicability'). However, these reductions in non-adherent testing were very small (0.5%, 0.4% and 0.3% respectively) and are unlikely to be clinically significant. Table 7.4 reports the model

coefficients with corresponding 95%CI and P values. Figure 7.4 presents the scatterplots for each model.

Table 7.2 Results from Models A.1 to A.6: Domains of AGREEII and non- adherent testing

AGREEII Domain	**Coefficient (95%CI)**	**P value**
Domain 1	0.1 (-0.5 to 0.7)	0.7
Domain 2	-0.5 (-0.7 to -0.2)	<0.01*
Domain 3	-0.4 (-0.6 to -0.1)	<0.01*
Domain 4	-0.3 (-0.9 to 0.3)	0.3
Domain 5	-0.3 (-0.6 to -0.01)	0.04*
Domain 6	-0.2 (-0.4 to 0.1)	0.2

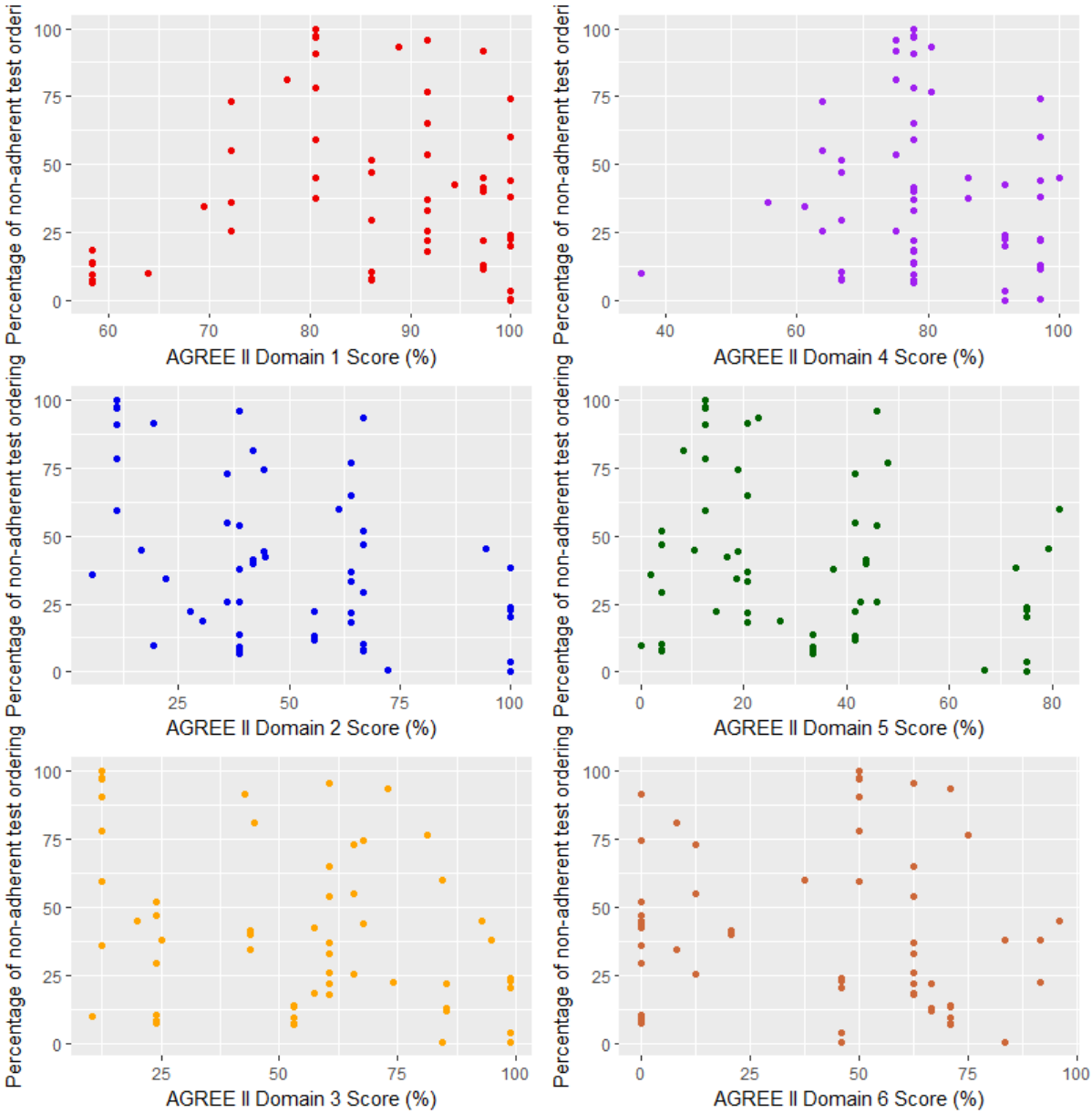


Figure 7.4 Distribution of scores for each Domain of the AGREEII tool

7.6.5 Association between guideline adherence and evidence quality (GRADE)

Model B showed a significant association between the quality of evidence (GRADE) and the percentage of non-adherent testing (Table 7.5). Guideline recommendations based on low and very low-quality evidence had 38% (95%CI 10% to 65%) and 24% (95%CI: 4.5% to 43%) more non-adherent testing, compared with guideline recommendations based on high quality evidence. Figure 7.5 presents the median percentage of non-adherent testing for each GRADE category.

Table 7.3 Results from Model B

GRADE category	**Coefficient (95%CI)**	**p value**
'High' Reference standard		
Low	37.7% (10.4% to 65.0%)	<0.01*
Very low	23.9% (4.5% to 43.3%)	0.02*

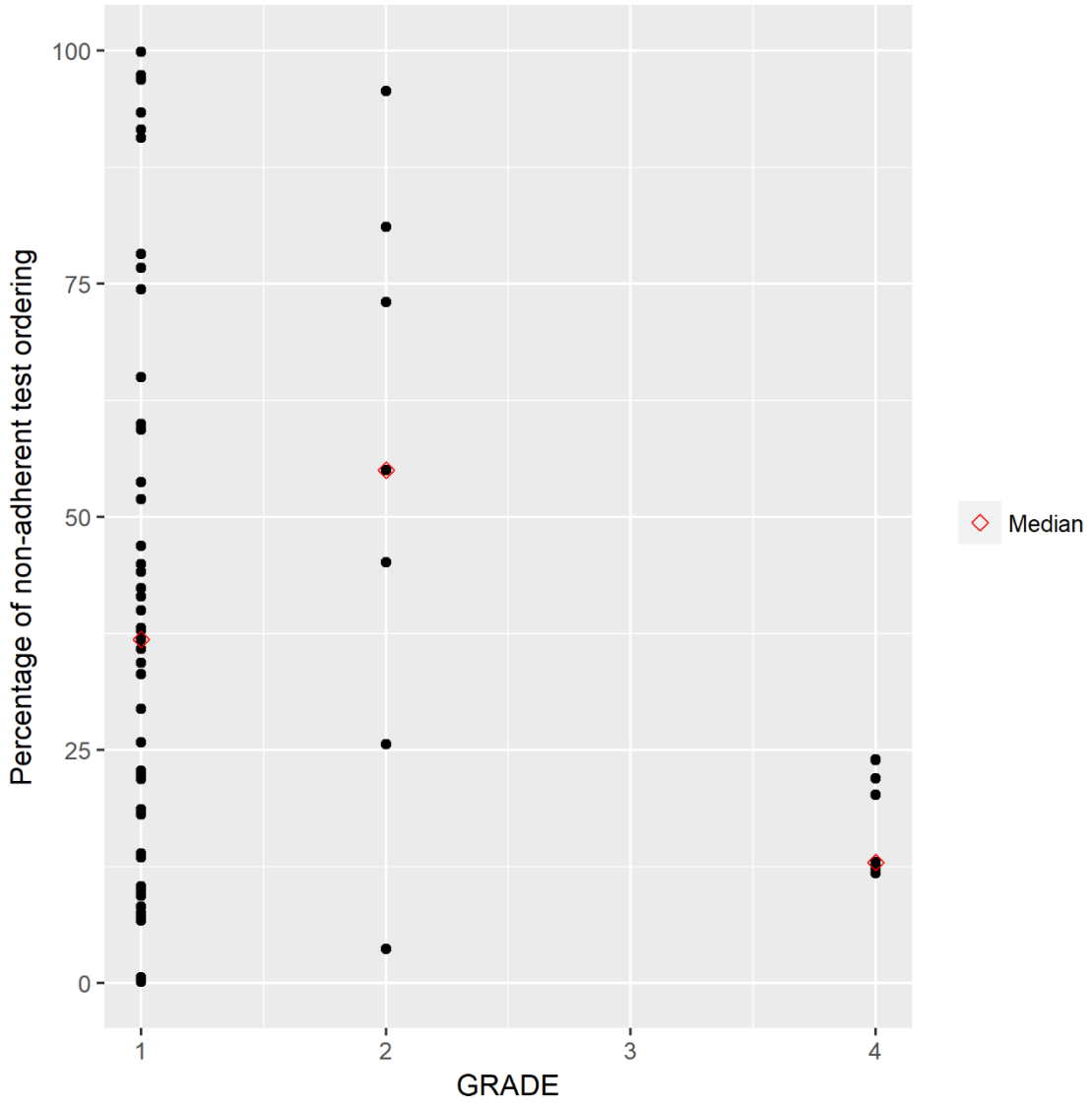


Figure 7.5 Scatter plot and median of each GRADE category and % guideline non-adherent test ordering (1 = Very low, 2 = Low, 3 = Moderate, 4 = High)

7.6.6 Sensitivity analysis

7.6.6.1 Low risk of bias

Percentage guideline non-adherence (0-100%) ~ f(guideline quality and reporting (Overall AGREE II Score, 1-7))

When only considering measures of non-adherence that were extracted from primary studies of low risk of bias, there was a significant decrease in non-adherence with an increase in AGREE II score

(coefficient: -6.0% (95%CI: -11.4% to -0.6%), P = 0.03). Although this association did not seem to follow a conventional linear relationship (figure 7.6).

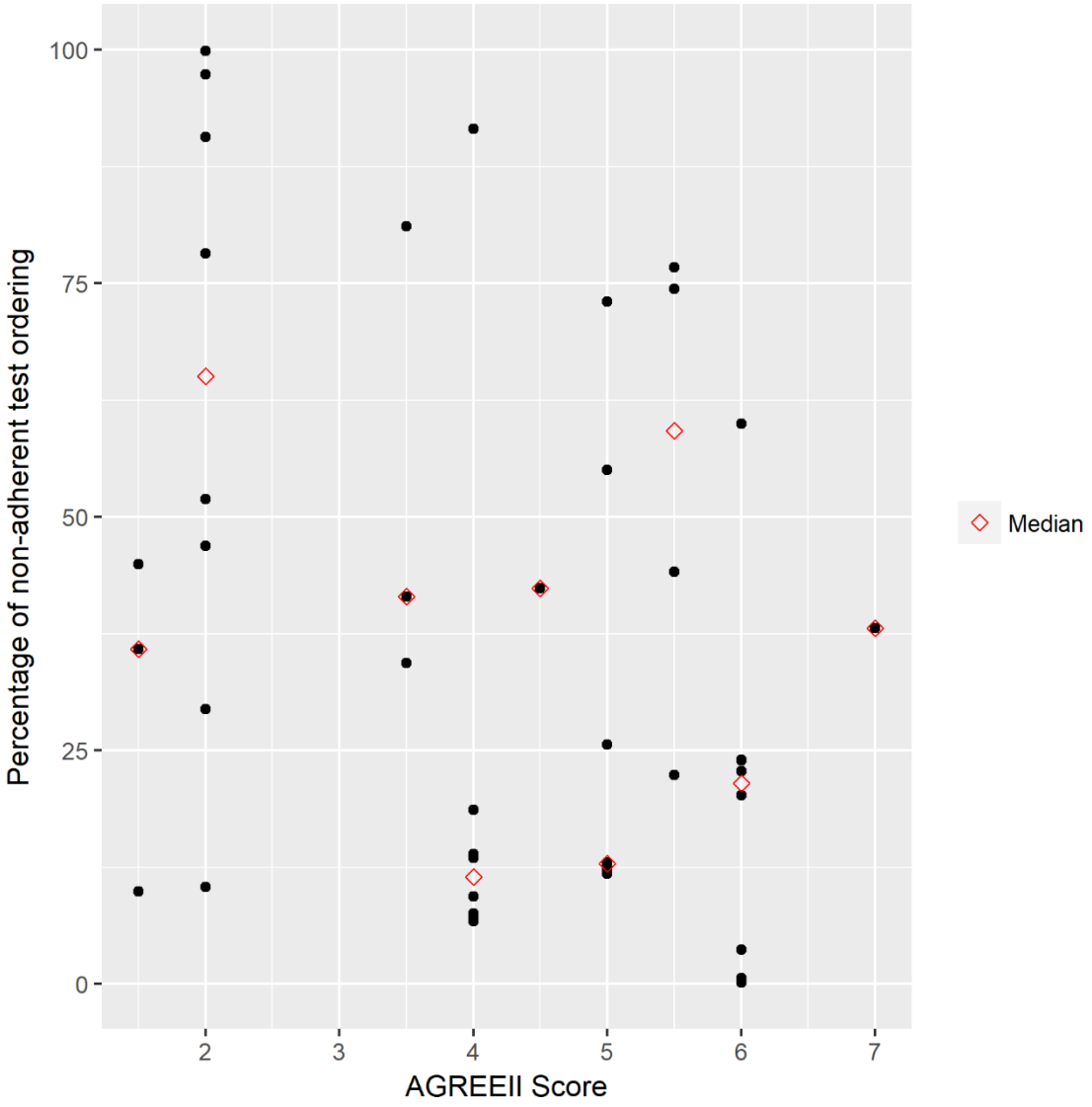


Figure 7.6 Low risk of bias studies: AGREE II scores and % guideline non-adherence

Association between AGREE II domains and guideline non-adherence

When only considering measures of non-adherence that were extracted from primary studies of low risk of bias, there was a significant decrease in non-adherence with an increase in AGREE II score for four of the six AGREE II domains (Table 7.6). Although these results were statistically significant, the absolute percentage decrease in non-adherence was very small; -0.3% to -0.5% respectively.

Table 7.4 Low risk of bias: Association between AGREE II Domains and guideline non-adherence

AGREE II Domain	Coefficient (95%CI)	P value
Domain 1	0.1% (-0.5% to 0.7%)	0.7
Domain 2	-0.5% (-0.8% to -0.2%)	<0.01*
Domain 3	-0.5% (-0.7% to -0.2%)	<0.01*
Domain 4	-0.4% (-0.1% to 1.2%)	0.2
Domain 5	-0.4% (-0.8% to -0.1%)	0.02*
Domain 6	-0.3% (-0.6% to -0.1%)	0.01*

Percentage guideline non-adherence (0-100%) ~ f(evidence quality (GRADE: Very low, low, moderate or high))

When only considering measures of non-adherence that were extracted from primary studies of low risk of bias, guideline recommendations based on low and very low-quality evidence had 32% (95%CI: 0.3% to 63%) and 24% (95%CI: 2.1% to 45%) more non-adherent testing, compared with guideline recommendations based on high-quality evidence.

7.6.6.2 Moderate risk of bias

Percentage guideline non-adherence (0-100%) ~ f(guideline quality and reporting (Overall AGREE II Score, 1-7))

When only considering measures of non-adherence that were extracted from primary studies of moderate risk of bias, there was a no significant decrease in non-adherent testing with an increase in AGREE II score (coefficient: 7.3% (95%CI: -3.8% to 18.5%), P = 0.2, figure 7.7).

Association between AGREE II domains and guideline non-adherence

When only considering measures of non-adherence that were extracted from primary studies of moderate risk of bias, there was no significant association between scores for any of the AGREE II

domains and guideline non-adherence. Table 7.7 reports the coefficients with their corresponding 95%CI and P values.

Table 7.5 Moderate risk of bias: Association between AGREE II Domains and guideline non-adherence

AGREE II Domain	Coefficient (95%CI)	P value
Domain 1	1.0% (-2.1% to 4.2%)	0.5
Domain 2	-0.5% (-1.3% to 0.4%)	0.3
Domain 3	0.3% (-0.3% to 1.0%)	0.3
Domain 4	0.3% (-1.2% to 1.9%)	0.7
Domain 5	0.6% (-0.1% to 1.2%)	0.1
Domain 6	0.4% (-0.1% to 0.8%)	0.1

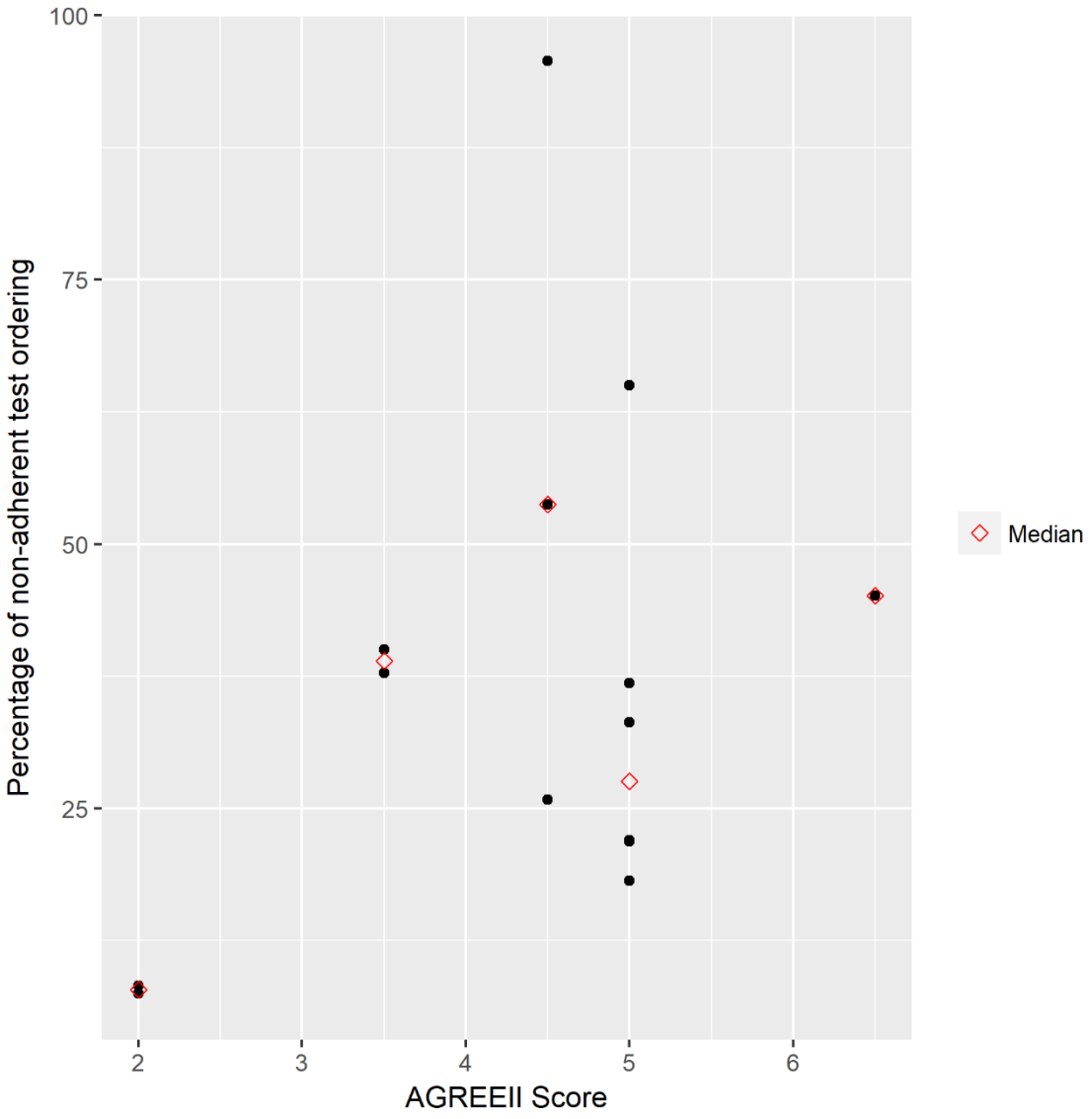


Figure 7.7 Moderate risk of bias: AGREE II scores and % guideline non-adherence

Percentage guideline non-adherence (0-100%) ~ f(evidence quality (GRADE: Very low, low, moderate or high))

When only considering measures of non-adherence that were extracted from primary studies of moderate risk of bias, there was no significant association between GRADE scores and guideline non-adherence. Table 7.8 reports the coefficients with their corresponding 95%CI and P values.

Table 7.6 Moderate risk of bias: Association between GRADE Score and guideline non-adherence

GRADE Score	Coefficient (95%CI)	P Value
'High Reference Standard		
Low	48.4% (-5.9% to 102.7%)	0.07
Very low	9.7% (-36.7% to 56.0%)	0.2

7.6.6.3 High risk of bias

Percentage guideline non-adherence (0-100%) ~ f(guideline quality and reporting (Overall AGREE II Score, 1-7))

When only considering measures of non-adherence that were extracted from primary studies of high risk of bias, there was a no significant decrease in non-adherent testing with an increase in AGREE II score (coefficient: 3.8% (95%CI: -99.3% to 107.0%), P = 0.7, figure S7.3). There were only three measures of non-adherence extracted from high risk studies.

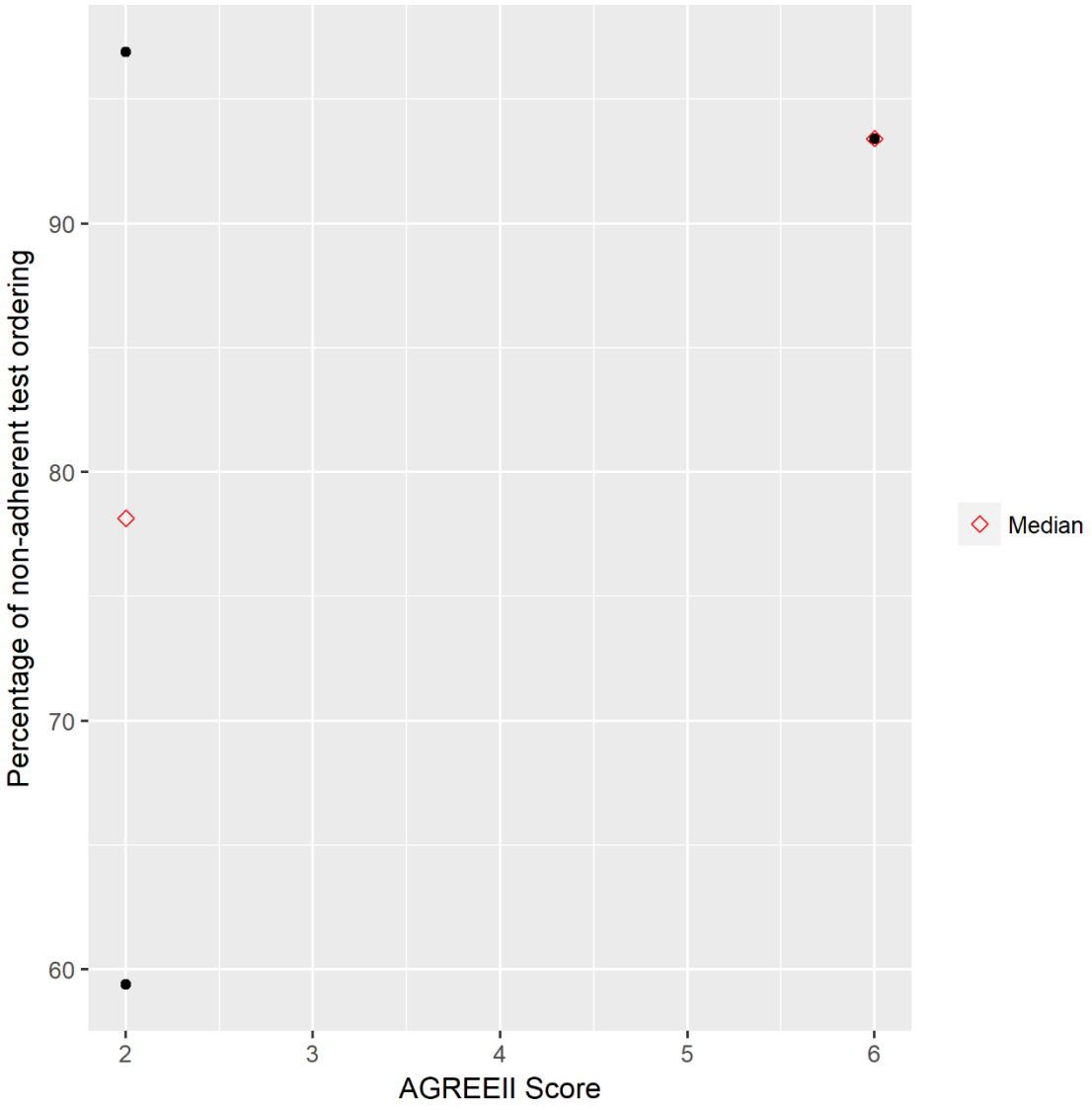


Figure 7.8 High risk of bias: AGREE II scores and % guideline non-adherence

Association between AGREE II domains and guideline non-adherence

When only considering measures of non-adherence that were extracted from primary studies of high risk of bias, there was no significant association between scores for any of the AGREE II domains and guideline non-adherence. Table S7.9 reports the coefficients with their corresponding 95%CI and P values.

Table 7.7 High risk of bias: Association between AGREE II domains and guideline non-adherence

AGREE II Domain	Coefficient (95%CI)	P value
Domain 1	1.8% (-47.9% to 51.6%)	0.5
Domain 2	0.3% (-7.2% to 7.7%)	0.7
Domain 3	0.3% (-6.6% to 7.1%)	0.7
Domain 4	5.5% (-144.0% to 155.0%)	0.7
Domain 5	1.5% (-38.1% to 41.1%)	0.7
Domain 6	0.8% (-19.1.1% to 20.5%)	0.7

Percentage guideline non-adherence (0-100%) ~ f(evidence quality (GRADE: Very low, low, moderate or high))

I could not perform an analysis exploring the association between GRADE score and non-adherence for measures of adherence from high risk of bias studies because all three measures had a GRADE score of very low.

7.6.7 Subgroup analysis

7.6.7.1 Underuse

Percentage guideline non-adherence (0-100%) ~ f(AGREE II Score (1-7))

For tests that had been underused, there was no significant association between guideline quality and reporting (AGREE II score) and the percentage of non-adherent testing for guidelines (P = 0.2). There was a non-significant decrease in non-adherent testing with an increase in AGREE II score (coefficient: -4.2% (95%CI: -11.2% to 2.9%)). Figure 7.9 shows the median percentage of non-adherent testing for each AGREE II score.

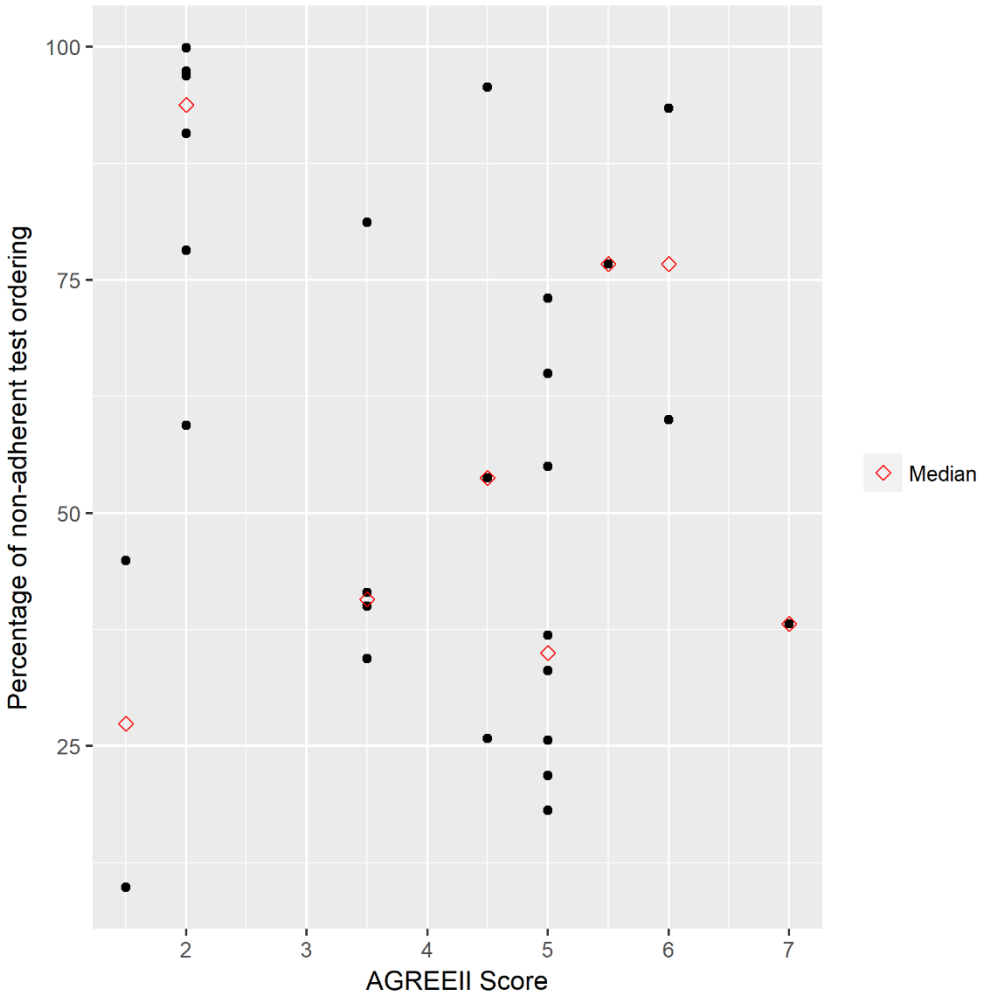


Figure 7.9 Underuse: Scatter plot and median of guideline quality and reporting (Overall AGREE II Score) and % guideline non-adherence

Association between AGREE II domains and guideline non-adherence

For tests that had been underused, there was no significant association between scores for any of the AGREE II domains and guideline non-adherence. Table 7.10 reports the coefficients with their corresponding 95%CI and P values.

Table 7.8 Underuse: Association between AGREE II Domains and guideline non-adherence

AGREE II Domain	Coefficient (95%CI)	P value
Domain 1	-0.2% (-1.3% to 1.0%)	0.8
Domain 2	-0.4% (-0.9% to 0.1%)	0.1

Domain 3	-0.3% (-0.7% to 0.1%)	0.2
Domain 4	0.7% (-0.3% to 1.7%)	0.2
Domain 5	-0.2% (-0.8% to 0.4%)	0.5
Domain 6	0.2% (-0.2% to 0.6%)	0.3

Association between GRADE score and guideline adherence

For tests that had been underused, there was no significant association between GRADE scores and guideline adherence. Table 7.11 reports the coefficients with their corresponding 95%CI and P values. Note that all the evidence pertaining to underuse had a GRADE score of either Low or Very low.

Table 7.9 Underuse: Association between GRADE score and guideline adherence

GRADE Score	**Coefficient (95%CI)**	**P Value**
'Low Reference Standard		
Very low	-10.8% (-39.5% to 17.8%)	0.4

7.6.7.2 Overuse

Percentage guideline non-adherence (0-100%) ~ f(AGREE II Score (1-7))

For tests that had been overused, there was no significant association between guideline quality and reporting (AGREE II score) and the percentage of non-adherent testing for guidelines (P = 0.6). There was a non-significant decrease in non-adherent testing with an increase in AGREE II score (coefficient: -1.4% (95%CI: -6.3% to 3.6%). Figure 7.10 shows the median percentage of non-adherent testing for each AGREE II score.

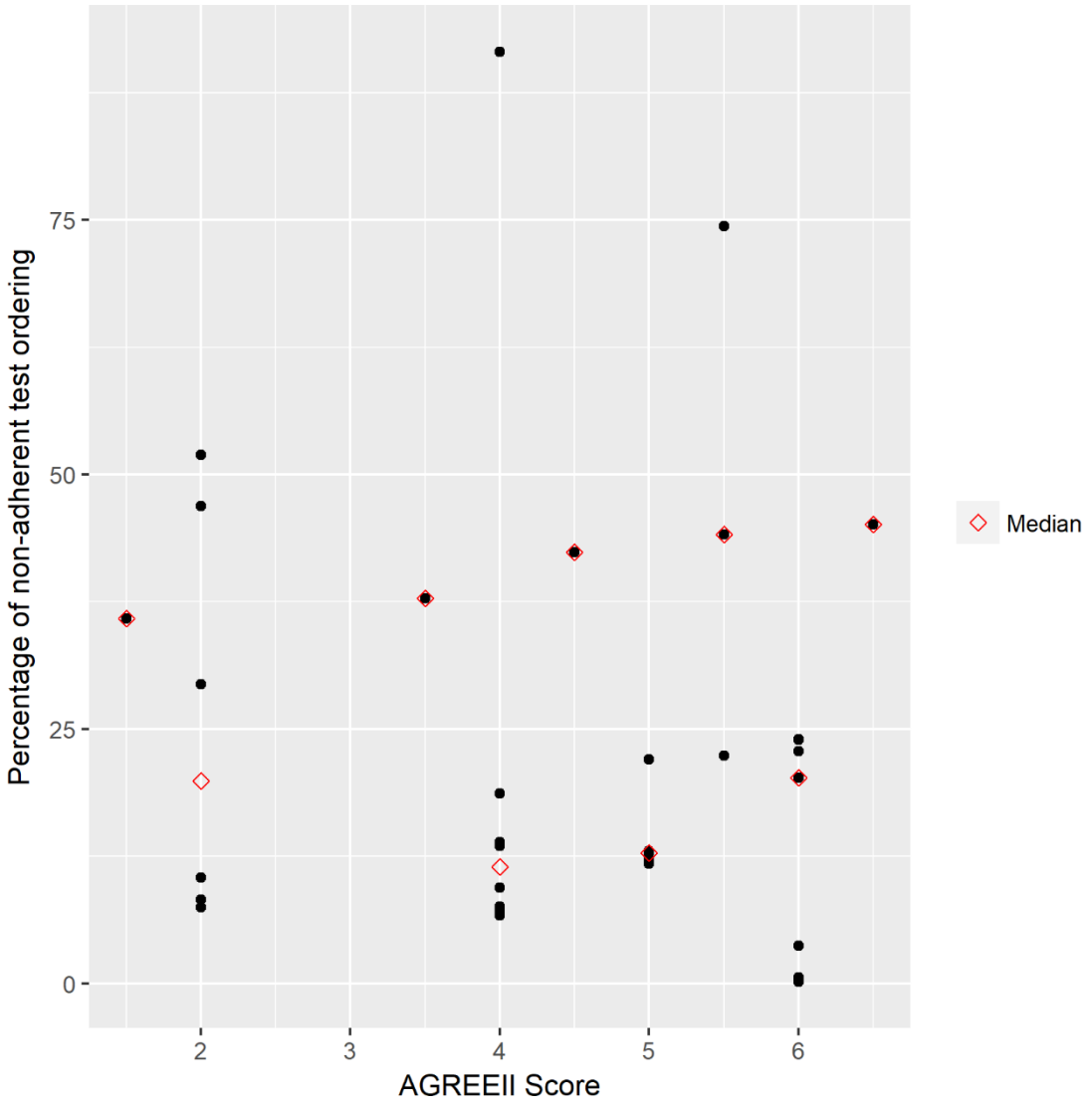


Figure 7.10 Overuse: Scatter plot and median of guideline quality and reporting (Overall AGREE II Score) and % guideline non-adherence

Association between AGREE II domains and guideline non-adherence

For tests that had been overuse, there was a significant association between domain 6 ("Editorial Independence") and guideline non-adherence. There was no significant association between the remaining AGREE II domains and guideline adherence. Table 7.12 reports the coefficients with their corresponding 95%CI and P values.

Table 7.10 Overuse: Association between AGREE II Domains and guideline non-adherence

AGREE II Domain	Coefficient (95%CI)	P value
Domain 1	0.003% (-0.1% to 0.7%)	0.8
Domain 2	-0.2% (-0.5% to 0.1%)	0.2
Domain 3	-0.2% (-0.4% to 0.1%)	0.1
Domain 4	-0.1% (-0.7% to 0.4%)	0.6
Domain 5	-0.2% (-0.5% to 0.03%)	0.1
Domain 6	-0.3% (-0.5% to -0.1%)	0.003*

Association between GRADE score and guideline adherence

For tests that had been overused, there was no significant association between GRADE scores and guideline non-adherence. Table S5.4 reports the coefficients with their corresponding 95%CI and P values.

Table 7.11 Overuse: Association between GRADE score and guideline non-adherence

GRADE Score	Coefficient (95%CI)	P Value
'High Reference Standard		
Low	7.9% (-24.7% to 40.5%)	0.6
Very low	9.7% (-6.2% to 25.7%)	0.2

7.7 Discussion

7.7.1 Summary of findings

I have assessed the robustness of guidelines in auditing the use of diagnostic tests. I determined the quality and reporting of guidelines and the quality of evidence underpinning recommendations. I then explored the relationship between the quality and reporting of guidelines (AGREE II score) and

guideline non-adherence and the relationship between the quality of evidence underpinning recommendations (GRADE) and guideline non-adherence.

There is an inconsistent relationship between the quality and reporting of guidelines (AGREE II) and guideline non-adherence. I did not find any relationship between overall AGREE II guideline score and the percentage of non-adherent testing. Although, this relationship became significant when only considering measures of non-adherence from low risk of bias studies. Furthermore, three of the domains of AGREE II were significantly associated with a decrease in non-adherent testing, these statistically significant results are unlikely to be clinically significant.

I found that most guidelines were based on very low-quality or low-quality evidence, and that guideline recommendations based on high-quality evidence had a lower percentage of non-adherent testing. Furthermore, recommendations based on high-quality evidence had significantly lower percentages of non-adherent testing.

Currently, the use of guidelines to directly measure overuse and underuse of diagnostic tests should be approached with caution. When a guideline has been developed adequately and its recommendations are based on high-quality evidence, guidelines may be appropriate tools with which to directly audit test use. However, my results suggest that this is rare.

4.7.2 Strengths and weaknesses in relation to other literature

To my knowledge, this is the first study exploring the associations between (a) the quality and reporting of a guideline (AGREE II) and non-adherent test use (Model A) and (b) the quality of evidence supporting a guideline recommendation and non-adherent testing (Model B). Other studies have assessed the quality and reporting of guidelines. A 2012 systematic review identified all studies that assessed the quality of guidelines using the AGREE tool (which preceded the AGREE II tool, but has the same domains).[26] The authors reported the mean score for each AGREE domain and found that the 'Scope and purpose' and 'Clarity and presentation' domains had the highest scores. They also reported that the 'Applicability' domain had the lowest score. My results are consistent with these

results. The use of the AGREE II tool to assess the quality and reporting of guidelines is a strength of my study. A 2013 systematic review concluded that of all the available clinical practice guideline appraisal tools, the AGREE II was the most comprehensive and appropriate English language tool to systematically assess guideline quality.[27]

My use of GRADE to assess the quality of the underlying evidence is also a strength. GRADE is used by over 100 organisations from 19 countries, including the WHO, the European Commission, the American College of Physicians, the UK's National Institute for Health and Care Excellence (NICE), UpToDate, and The BMJ. Few studies have audited the quality of the evidence supporting guideline recommendations. A 2016 study[28] systematically appraised all WHO guidelines between 2007 and 2012 and found, as we have, that most recommendations were based on poor-quality evidence. The authors reported that 'over 50% of WHO strong recommendations are based on low or very low [evidence]'.[28]

My study has several limitations. I found only two-thirds of the relevant guidelines. It is plausible that these missing data would change my results quantitatively, but it is unlikely that extra data would change our conclusion: that most guidelines are based on poor-quality evidence, guideline quality and reporting have an unclear effect on adherence, but the quality of evidence supporting guideline recommends is associated with adherence.

For some of the guideline recommendations it was unclear if evidence was cited to support the recommendation. It was not uncommon for guidelines to report 'research suggests', but then not provide a reference. In these uncertain cases I emailed the professional bodies, seeking clarification. If the guideline organisation did not respond or could not provide the primary reference supporting the recommendation we considered the recommendation to be 'Expert Opinion'. It is possible that I misclassified some guideline recommendations as 'Expert Opinion'.

Further, as mentioned in the methods, in some instances, there were multiple indications for test use within one guideline recommendation. This was problematic because I had only one measure of non-adherence diagnostic test ordering. In these situations, it would be ideal to have a measure of non-

adherence for each specific of the indications within the guideline recommendation, but I was restricted by the primary studies within chapter 6, which only provided one measure of non-adherence. This situation was only relevant for 8 of the 62 (13%) measures of non-adherence.

It is also important to note that I only assessed guidelines that had been used in the primary studies in Chapter 6. Wider assessment of all currently available guidelines may provide more statistical power to evaluate the quality and reporting of guidelines and the quality of evidence underpinning the recommendations. However, this approach would not allow us to achieve our third and fourth objectives; to examine the association between the quality and reporting of guidelines and non-adherence and the association between evidence quality and non-adherence.

Furthermore, chapter 6 captured all primary studies that measured the adherence of *any* primary care diagnostic test against a relevant national or international guideline. However, not all diagnostic tests have been studied in the primary literature. Therefore, it was not possible to include all diagnostic tests available to clinicians. Similarly, I have not examined all diagnostic test guidelines available. It is plausible that if the data were available for *all* tests (and their respective guidelines) my results may differ.

Although not strictly a methodological limitation, it is important to consider that guideline recommendations may not be applicable to all patients in all clinical situations. In this study, I did not determine nor state the appropriate rate of guideline adherence, nor do I explore when non-adherence to a recommendation may be appropriate – for example, if a clinician and patient reach a shared decision to not investigate a symptom due to a patient's advanced age. It remains contentious, on a population-level, what rate of guideline adherence is acceptable, and, in a sense, my study provides data to help explore this issue. I do not imply that clinicians and patients should always follow guidelines; there will always be times when it is appropriate to deviate from recommendations. Medical decisions should consider the best available evidence, clinical guidance, clinician expertise and experience and patient values. I present data that shows that diagnostic test recommendations

based on high-quality evidence have greater rates of adherence, compared to those based on low or very-low quality of evidence. We feel this novel finding will be of value to guideline developers.

7.7.3 Implications

7.7.3.1 Implications for patients

Domain II of the AGREE II tool ('Stakeholder Involvement') directly addresses the involvement of patients and the public in the development of guidelines. This Domain had the second lowest median score; guidelines scored on average 44% of what is ideally required in this section of a gold-standard guideline. Greater involvement of patients in guidelines can improve the quality of guidelines.[29] Further patient involvement can also help guidelines transition from didactic documents to ones that encourage shared decision making,[30] which may also lead to greater adherence to recommendations.

7.7.3.2 Implications for clinicians

Our results have highlighted the caution that clinicians should exert when following clinical practice guidelines. Most recommendations I located were based on no evidence at all and I noted consistent flaws in guidelines, most significantly lack of patient involvement in guideline construction and paucity of advice about how to implement guidelines in routine clinical practice (Domain 5).

7.7.3.3 Implications for policy makers

The implications for guideline makers are clear. Guideline recommendations based on high-quality evidence are more likely to be adhered to. When high-quality evidence is available, recommendations should be based on the available evidence. Nevertheless, there remains clinical questions where there is a paucity of high-quality evidence. In these scenarios, it is not possible for guideline developers to base their recommendations on high quality evidence. When this is the case, it is important for the strength of the recommendation to reflect the underlying evidence, i.e. it is less likely to be appropriate for recommendations based on very-low quality evidence to warrant a strong

recommendation. Unfortunately, in the past many guideline recommendations with a "strong" strength of recommendation have been based on low or very low evidence [28].

There are further implications for policy makers: 1. Guideline developers can highlight evidence gaps. If they develop a guideline and only find very-low or low-quality evidence to support their recommendation, this information should be publicised, especially to academic centres, who can fill these gaps. 2. Policy makers may want to consider targeting performance indicators around guideline recommendations that are supported by high-quality evidence and have a strong strength of recommendation. Conversely, where only very-low or low evidence exists, targeted performance indicators are less likely to be appropriate.

7.7.3.4 Implications for researchers

One of the reasons clinical practice guidelines were established to mitigate geographical variations in practice. However, there is still substantial geographical variations in diagnostic test ordering.[5] Varying adherence to guidelines is a probable driver of unwarranted variation, and research efforts to mitigate non-adherence would be advantageous, specifically, research efforts directed at interventions to improve adherence to guideline recommendations based on high-quality evidence.

7.7.4 How this chapter fits within the wider thesis

My thesis attempts to develop methods to identify inappropriate ordering of tests. Chapters 3-5 explore indirect indicators of potential inappropriate use. Chapter 6 directly quantifies inappropriate overtesting and undertesting, using guidelines as the metric against which appropriateness is measured. This chapter directly leads on from Chapter 6. An accepted limitation of using guidelines as the assessment metric is their varying quality.[4,31] In this chapter, I tested the robustness of guidelines as tools to directly audit the ordering of diagnostic tests.

7.8 Conclusion

Guidelines are often used to directly quantify over and undertesting.[5,32] I have shown that many diagnostic test guideline recommendations are based on evidence of low or very low quality. There is

also much heterogeneity in the quality and reporting of guidelines. Thus, to use guidelines as tools to audit the appropriateness of test ordering, recommendations need to be supported by better evidence and guideline development needs to be consistently of a higher quality. Currently, few guidelines meet this requirement.

Our results do, however, present promising evidence that guideline recommendations based on high-quality evidence are adhered to and thus provide guideline developers further incentives to appropriately consider evidence in their recommendations.

7.9 Chapter Summary

- My thesis attempts to identify inappropriate ordering of diagnostic tests. It proposes indirect indicators of potential inappropriate use and uses guidelines as a tool to directly measure inappropriate overuse and underuse (Chapter 6). Chapter 7 assesses the robustness of guidelines as a tool to directly audit test use.
- The quality and reporting of guidelines used to measure non-adherent diagnostic test ordering from Chapter 6 (SR) were assessed using the AGREE II tool. The quality of evidence supporting guidelines recommendations was then assessed using GRADE.
- I then assessed the relationship between guideline non-adherence and the quality and reporting of guidelines (AGREE II), as well as the relationship between guideline non-adherence and evidence quality (GRADE).
- My results show that most guidelines are based on evidence of low or very low quality and there is much heterogeneity in their quality and reporting.
- Guideline recommendations based on high-quality evidence have a significantly lower percentage of non-adherent testing.
- To use guidelines as tools to directly audit the appropriateness of test ordering, recommendations need to be supported by better evidence and guideline quality and reporting needs to be consistently of a higher quality. Currently, few guidelines meet this requirement.

References

1 Chalmers K, Pearson S, Elshaug AG. Quantifying low-value care: a patient- versus service-centric lens. *BMJ Qual Saf* 2017; **26**: 855–8.

2 Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. *JAMA* 1999; **281**: 1900–5.

3 Grilli R, Magrini N, Penna A, *et al.* Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet (London, England)* 2000; **355**: 103–6.

4 Lenzer J. Why we can't trust clinical guidelines. *BMJ* 2013; **346**.

5 O'Sullivan JW, Albasri A, Nicholson B, *et al.* Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open* 2018; **8**: e018557.

6 Brouwers MC, Kho ME, Browman GP, *et al.* AGREE II: Advancing guideline development, reporting and evaluation in health care. *J Clin Epidemiol* 2010; **63**: 1308–11.

7 Mafi JN, McCarthy EP, Davis RB, Landon BE. Worsening Trends in the Management and Treatment of Back Pain. *JAMA Intern Med* 2013; **173**: 1573.

8 Chou R, Qaseem A, Snow V, *et al.* Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. *Ann Intern Med* 2007; **147**: 478–91.

9 Kerry S, Hilton S, Dundas D, Rink E, Oakeshott P. Radiography for low back pain: A randomised controlled trial and observational study in primary care. *Br J Gen Pract* 2002; **52**: 469–74.

10 Kendrick D, Fielding K, Bentley E, Kerslake R, Miller P, Pringle M. Radiography of the lumbar spine in primary care patients with low back pain: randomised controlled trial. *BMJ* 2001; **322**: 400–5.

11 Guyatt GH, Oxman AD, Vist GE, *et al.* GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; **9**: 8–11.

12 Guyatt GH, Oxman AD, Vist G, *et al.* GRADE guidelines: 4. Rating the quality of evidence - Study limitations (risk of bias). *J Clin Epidemiol* 2011; **64**: 407–15.

13 Guyatt GH, Oxman AD, Kunz R, *et al.* GRADE guidelines: 7. Rating the quality of evidence - Inconsistency. *J Clin Epidemiol* 2011; **64**: 1294–302.

14 Guyatt GH, Oxman AD, Kunz R, *et al.* GRADE guidelines: 8. Rating the quality of evidence - Indirectness. *J Clin Epidemiol* 2011; **64**: 1303–10.

15 Guyatt GH, Oxman AD, Kunz R, *et al.* GRADE guidelines 6. Rating the quality of evidence - Imprecision. *J Clin Epidemiol* 2011; **64**: 1283–93.

16 Rogozinska E, Khan K. Grading evidence from test accuracy studies: what makes it challenging compared with the grading of effectiveness studies? *Evid Based Med* 2017; **22**: 81–4.

17 Higgins JPT, Altman DG, Gotzsche PC, *et al.* The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Bmj* 2011; **343**: d5928–d5928.

18 Schünemann HJ, Oxman AD, Brozek J, *et al.* GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008; **336**: 0–b.

19 Schünemann HJ, Mustafa R, Brozek J, *et al.* GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. *J Clin Epidemiol* 2016; **76**: 89–98.

20 Mustafa RA, Santesso N, Khatib R, *et al.* Systematic reviews and meta-analyses of the accuracy of HPV tests, visual inspection with acetic acid, cytology, and colposcopy. *Int J Gynecol Obstet* 2016; **132**: 259–65.

21 Shea BJ, Grimshaw JM, Wells GA, *et al.* Development of AMSTAR: a measurement tool to

assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007; **7**: 10.

22 Whiting PF, Rutjes AWS, Westwood ME, *et al.* QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med* 2011; **155**: 529–36.

23 Cochrane Collaboration. Tool to assess risk of bias in cohort studies. *Cochrane Methods* 2013; : 1–4.

24 Shaheen NJ, Weinberg DS, Denberg TD, Chou R, Qaseem A, Shekelle P. Upper endoscopy for gastroesophageal reflux disease: best practice advice from the clinical guidelines committee of the American College of Physicians. *Ann Intern Med* 2012; **157**: 808–16.

25 Hoy D, Brooks P, Woolf A, *et al.* Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement. *J Clin Epidemiol* 2012; **65**: 934–9.

26 Alonso-Coello P, Irfan A, Solà I, *et al.* The quality of clinical practice guidelines over the last two decades: A systematic review of guideline appraisal studies. *Qual Saf Heal Care* 2010; **19**: 1–8.

27 Siering U, Eikermann M, Hausner E, Hoffmann-Eßer W, Neugebauer EA. Appraisal tools for clinical practice guidelines: A systematic review. *PLoS One* 2013; **8**. DOI:10.1371/journal.pone.0082915.

28 Alexander PE, Brito JP, Neumann I, *et al.* World Health Organization strong recommendations based on low-quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. *J Clin Epidemiol* 2016; **72**: 98–106.

29 Diaz del Campo P, Gracia J, Blasco JA, Andradas E. A strategy for patient involvement in clinical practice guidelines: methodological approaches. *BMJ Qual Saf* 2011; **20**: 779–84.

30 Berger ZD, Brito JP, Ospina NS, *et al.* Patient centred diagnosis: sharing diagnostic decisions with patients in clinical practice. *Bmj* 2017; **4218**: j4218.

31 Burgers JS, Fervers B, Haugh M, *et al.* International Assessment of the Quality of Clinical

Practice Guidelines in Oncology Using the Appraisal of Guidelines and Research and Evaluation Instrument. *J Clin Oncol* 2004; **22**: 2000–7.

32 Zhi M, Ding EL, Theisen-Toupal J, *et al.* The Landscape of Inappropriate Laboratory Testing: A 15-Year Meta-Analysis. *PLoS One* 2013; **8**: e78962.

Chapter Eight: Discussion

8.1 Chapter Aim

The aim of this chapter is to summarise the thesis and its findings, limitations and implications; including a proposed approach to identify potentially inappropriate use of tests.

8.2 Introduction

Globally, the sustainability of healthcare systems is under threat. In the UK, rising expenditure, coupled with insufficient funding, has left the NHS in a precarious financial state.[1] Currently, the NHS is not projected to meet the targeted £22 billion in efficiency savings.[2] The identification of wasteful spending remains a priority of the Department of Health and Clinical Commissioning Groups (CCGs).

8.3 Summary of findings

In chapter one I have given an overview of my thesis, its structure and my achievements. In chapter 2, I described the background to the wider topic of healthcare sustainability. Specifically: waste in healthcare, the role and importance of tests in practice, the cost of using tests, drivers of inappropriate use, previous efforts to quantify unnecessary use of resources and my personal and professional rationale for pursuing this research.

My aim in chapter 3 was to determine which tests had the greatest change in use between 2000 and 2015. To achieve this, I chose 44 commonly used tests (28 laboratory, 11 imaging, and five miscellaneous tests). I found that, MRI knee scans had the highest average annual increase over 16 years, followed by vitamin D tests and MRI brain scans. Vitamin D, MRI knee, and CT pelvis had the greatest increases after QOF. Vaginal swab was the only test in which the use fell significantly over the study period.

In chapter 4 I reported the results on between-practice variation in test use. My aim in this chapter was to identify tests with the greatest between-practice variation in use. I studied the same 44 tests used in

chapter 3. The tests with the highest variation were non-illicit drug test, pap smear and pelvic CT. I also identified tests that had both higher than average between-practice variation and a higher than average rate of use. Seven tests met this classification: Clotting, Vitamin D, Urine Albumin, Prostate Specific Antigen (PSA), urine MCS, Bone profile, and C-reactive Protein (CRP)) tests.

In chapter 5 I identified general practices that ordered a lower proportion of tests that returned an abnormal result (i.e. outside the reference range), compared with their colleagues. I explored CRP and Erythrocyte Sedimentation Rate (ESR) tests ordered across Oxfordshire NHS General Practices. I analysed 143,745 CRP and 30,758 ESR requests from 69 practices and identified twelve (17%) and seven (10%) practices that ordered a significantly lower proportion of abnormal CRP and ESR tests respectively.

In chapter 6 my aim was to explore primary care physicians' non-adherence with diagnostic test guidelines. I conducted a systematic review, which included primary studies that quantified either underuse, failure to order a test when indicated, or overuse, ordering a test when not indicated (or both). 17 tests were underused >50% of the time. Of these, echocardiography (n=4 measures) was consistently underused (between 54% and 89%, n=4). There was large variation in the rate of inappropriate underuse of pulmonary function tests (38%–78%, n=8). Eleven tests were inappropriately overused >50% of the time. Echocardiography was consistently overused (77%–92%), whereas inappropriate overuse of urinary cultures, upper endoscopy and colonoscopy varied widely, from 36% to 77% (n=3), 10% to 54% (n=10) and 8% to 52% (n=2), respectively.

In chapter 7 I aimed to explore the robustness of guidelines to directly audit the appropriateness of test use. I assessed guidelines used to determine the appropriateness of test use in the systematic review reported in chapter 6. I also assessed the quality of evidence supporting these guideline recommendations. We (Ali Albasri also independently did this) used the Appraisal of Guidelines for Research & Evaluation (AGREE) II tool to assess the quality and reporting of guidelines, we also used Grading of Recommendations Assessment, Development, and Evaluation (GRADE) to examine

the quality of evidence underpinning these recommendations. Finally, I constructed linear regression models to determine the association between guideline non-adherence (data from Chapter Six) and a) the guideline AGREE II score and b) the GRADE score. The AGREE II scores for the guidelines were varied but largely suboptimal (median 4.5 (IQR: 3.5-5)). Most (n=45, 73%) of the evidence underpinning guideline recommendations was graded as very low, with seven graded as low (11%), and ten graded as high (16%) (none was graded as moderate). There was no significant association between the AGREE II score and non-adherent testing (p = 0.09). However, this relationship became significant when only considering measures of non-adherence from low risk of bias studies. Furthermore, there was a significant association between the GRADE score and non-adherence. Guideline recommendations based on low-quality and very low-quality evidence had 38% (95%CI 10% to 65%, P = <0.01) and 24% (95%CI: 4.5% to 43%, P = 0.02) more non-adherence, compared with recommendations based on high quality evidence.

8.4 Strengths and limitations in relation to previous research

Previous research has quantified potentially unnecessary care directly or indirectly. Indirect measures take a population-level approach and typically describe 'unexpected variation' in healthcare resource use; variation that persists despite adjustment for differences in patients or population characteristics between regions or over time.[3] Direct assessment of over and underuse of healthcare resources requires an examination of individual patient data (IPD). Mostly, this involves a comparison of IPD with a high-quality 'appropriateness criteria'; guidelines have been the most commonly used 'appropriateness criteria'.[4–7]

In my thesis, I have examined three indirect, population-level metrics for test under and overuse. I have identified the primary care tests subject to the greatest temporal increases in use, as well as identified the test subject to the greatest between-practice variation and identified general practices that ordered a lower proportion of ESR and CRP tests that return an abnormal result. I am the first to examine the temporal changes in laboratory, imaging and other, miscellaneous tests in UK primary care. A previous study[8] assessed the change in 29 laboratory tests from UK primary care from 2005 to

2009 in a random sample of 100,000 patients. My study looked at more tests (44), all test types (laboratory, imaging and other), over a longer period (2000 to 2015 vs. 2005 to 2009) and in an unrestricted patient sample (I included all tests that had occurred over our time period, the other authors looked in a random sample restricted to 100,000 patients). To my knowledge, no other studies have identified primary care tests subject to the greatest temporal change in use.

The same previous study[8] ranked 29 laboratory tests by their between-practice variability. This analysis was similarly restricted in its type (laboratory only) and number of tests, but also used a different method to quantify between-practice variation. This study ranked tests by the standard deviation in their rate of use; this is not the most appropriate reflection of the variation because standard deviation is affected by sample size (the frequency of test use). Therefore, tests that are most commonly ordered are more likely to have lower standard deviations. I ranked the between-practice variation of test use by their co-efficient of variation (CoV); which accounts for the confounding effect of sample size. The use of CoV is recommended by The King's Fund.[9]

To my knowledge, I am the first to identify general practices that order a significantly lower proportion of abnormal tests (tests that return an abnormal result). However, my methods were adapted from Spiegelhalter[10]. Spiegelhalter describes the use of funnel plots to compare institutional performance; he suggests the construction of control limits two and three standard deviations from a target outcome – typically the sample mean of the outcome. Although his method has never been applied to examine test use, it has been used to examine performance between surgeons,[11] emergency readmissions to hospital,[12] and mortality differences between general practices.[13]

No study, to my knowledge, has quantified overuse and underuse of primary care diagnostic tests against guidelines. Zhi et al[5] quantified the mean rates of overuse and underuse of laboratory tests in secondary care and focused on quantifying an overall rate of overuse and underuse. They estimated that overuse and underuse of laboratory tests was around 21% and 45%, respectively. I chose not to quantify an overall rate of overuse and underuse because I felt combining this data would be inappropriate (I have further described this in *Chapter 6)*. A 2015 systematic review examined all

electronic health record-based interventions for improving the appropriateness of diagnostic imaging.[41] This study also included primary studies that used guidelines to quantify appropriateness. It should be noted that these studies – studies that examined the appropriateness of healthcare resource use against a guideline – often use the term "low-value care"[42] to refer to care not in line with guidelines. I have avoided this term in this thesis. While the term "low-value care" may be considered by many to be adequate, I find this term ambiguous. I therefore preferred to be more explicit and used the phrase: a direct assessment of the appropriateness of test use against guidelines.

A 2012 study presented an approach to identify health care services of low-value.[40] This review set out to identify services (tests and treatments) that were currently recommended (approved for reimbursement on the Australian Medicare Benefits Schedule), but their use was not supported by evidence. The methods they employed were an extensive and systematic literature search for evidence that suggested a medical practice was not efficacious. They searched PubMed and targeted databases such as The Cochrane Library and NICE Do Not Do recommendations. They also used opportunistic sampling to identify other literature. They identified tests and treatments where the underlying evidence did not support Australian reimbursement practice, i.e. they identified recommendations where evidence suggested that the practice should not be performed. My thesis, conversely, identified tests and clinicians who may be ordering tests when they are not currently recommended.

Furthermore, I was the first to assess the robustness of diagnostic test guidelines to audit test use. A strength of this study is my use of the most valid guideline appraisal tool (AGREE II) and the most widely accepted and used method to assess the quality of evidence (GRADE).

My thesis is not without limitations. There are limitations of my individual research studies, I have discussed these extensively within the specific results chapters (*Chapter 3 to 7),* I will discuss the most pertinent points here. There are also broad limitations of the thesis to discuss.

It is important to discuss the limitations of indirect, population-level markers of potential waste. These studies (*Chapter 3-5)* highlight variation not accounted for by some differences in practice populations; variation that may be worth investigating further. In isolation, temporal change in use

and marked between-practice variation do not necessarily imply inappropriate use, nor does the ordering of tests the return a lower proportion of abnormal results. For example, there may be appropriate reasons that a test was subject to substantial increases in use and there may also be legitimate reasons for between-practice variation. Although adjustment for differences in patient populations can help control for these differences, it is plausible that some warranted between-practice variation may persist. Similarly, general practices may have legitimate reasons for ordering a higher or lower proportion of abnormal tests. I have highlighted variation, with no clear cause, that may benefit from further investigation.

Chapter 6 is limited by the use of guidelines as a criterion to determine the appropriateness of test use directly. Guidelines have been criticized for varying quality, particularly, biases in their development[14,15] and lack or poor quality evidence supporting their recommendations.[16] These limitations jeopardise the use of guidelines for directly auditing the appropriateness of test use.

In Chapter 7, I acknowledged this as a limitation and assessed the robustness of guidelines for auditing test use. I showed that of the diagnostic guidelines that I examined, most were based on low or very low-quality evidence. As such, not all guidelines appear to be optimal tools for auditing the appropriateness of tests. I highlight the need for better evidence to be produced in *8.4.6 Next steps.*

Another limitation of chapter 6 is data quality in primary care patient records. There are reasonable concerns about the consistency of patient records; for example, records may not contain sufficient detail to confidently class the use of a test as inappropriate, or this level of detail may not be consistent across many patient records. It is unlikely that all primary care records will contain data of adequate quality. Nevertheless, the RAND corporation, in the USA, has measured overuse retrospectively using primary care records.[17–19] Therefore, it appears that in at least one setting, primary care records are considered of sufficient quality to confidently quantify overuse. Furthermore, several UK studies have also attempted to retrospectively quantify overtesting, using a similar method (assessing primary care health records, some using the CPRD).[20,21] However, these studies seem to have focused on tests for which there are clear indications and contraindications; it remains unlikely

that the CPRD (or any other electronic primary care health records database) will have sufficiently high-quality data to retrospectively quantify overtesting across a large variety of tests. As electronic health records continue to improve, this should become less of a limitation. In the meantime, a prospective, rather than retrospective, audit of the appropriateness of test use may be a solution.

More broadly, most of my research was conducted in one setting: the UK. Chapters 3,4 and 5 use national and local (Oxfordshire) UK General Practice data. It was not possible to conduct research in other settings due to data and resource requirements. It could be argued that this is not a limitation, more a point concerning the external validity of the research. Regardless of the nomenclature, it must be acknowledged that most of my research is restricted to one country. Similarly, all my research concerns primary care. While the methods I employ could probably be used to assess secondary care test use, the scope of my research was restricted to primary care.

Furthermore, I did not conduct research exploring *why* variation in primary care test use exists. I did not determine the reasons why there has been such an increase in test use over time, nor why some tests increased at a greater rate than others. Similarly, I did not explore explanations for the noted between-practice variation in test use. It was also beyond the scope of Chapter 5 to determine why some general practices ordered a lower proportion of tests that returned an abnormal result. For all of these chapters I did adjust for demographic differences between practices, but future research will likely require a mixed methods approach to address these outstanding issues.

Furthermore, I did not seek to include patients in all aspects of my research. Chapters 3 and 4 benefited from a dedicated Patient and Public Involvement (PPI) group that assisted at the protocol, analysis and interpretation stage of the research. These chapters undoubtedly benefited from their input. The other studies did not have the funding to support dedicated patient involvement. It would be of great benefit to involve patients in the next stages of this research; I am sure they will help immensely in explaining some of the variation in test use I have noted in my research.

For each of my reported studies, I have described the strengths in their respective chapters. The broad strengths of my thesis include an array of quantitative approaches, the assessment of many aspects of

test use (temporal increase, between-practice variation, frequency of use and variation in the proportion of tests that return an abnormal result), the utilisation of both national and local data and diverse methodological techniques (systematic reviews, meta-epidemiological studies and big data analyses).

8.5 Implications for policy and practice

8.5.1 A potential method to identify inappropriate diagnostic test use

At the start of my DPhil, I aimed to assess the use of tests in primary care. As I familiarised myself with the background literature it became apparent that metrics to identify potential waste in healthcare were broadly categorised as "direct" or "indirect" measures.[22–24] As described in *Chapter 2*, indirect measures take a population-level approach and commonly report "unexpected variation", whereas direct measures consider the individual patient.

As I conducted my research, I explored indirect markers of waste: temporal change in test use, between-practice variation in use and variation in the proportion of tests that returned an abnormal result. I also was the first to directly quantify the appropriateness of primary care test use against guidelines (it had previously been done for secondary care laboratory tests). While doing this research, it became apparent that I could combine my indirect and direct metrics into a 4-step approach to identify potentially inappropriate test use.

The combination of my various metrics into one approach seemed logical, because it would allow me to discuss practical steps that policy makers, doctors and researchers could follow. A structured process to try and identify waste is important, particularly in the UK, given the financial pressures currently on the NHS and the potential harms of unnecessary test use. To my knowledge, no one has published a similar strategy, nor does there seem to be policy in place to help general practices and Clinical Commissioning Groups (CCG) mitigate inappropriate test use.

In the following section I will explain four steps that may be able to be used to identify potentially inappropriate test use. I will explain each step, the implications of my proposed approach, its

limitations and the future research required. Originally, these four steps were designed to identify both underuse and overuse of tests, although the four steps I ultimately propose (below) are only intended to identify overuse – I elaborate on this point in the limitations section *8.4.2*. Briefly, the four-steps are (figure 8.1):

1. The identification of tests with the greatest combination of temporal increase, between-practice variation, and frequency of use. In my thesis, Chapters 3 and 4 explain the methods to achieve this step.
2. A discussion between stakeholders to finalise the tests they want to focus on from the selection captured in step 1. This step is not derived from any of the chapters of this thesis.
3. The identification of general practices that order a lower proportion of tests that return an abnormal result, compared to their colleagues. Chapter 5 explains the methods to achieve this step.
4. An audit of the appropriateness of test use by comparing individual patient data with relevant national or international guidelines. Chapter 6 explains the methods to achieve this step.

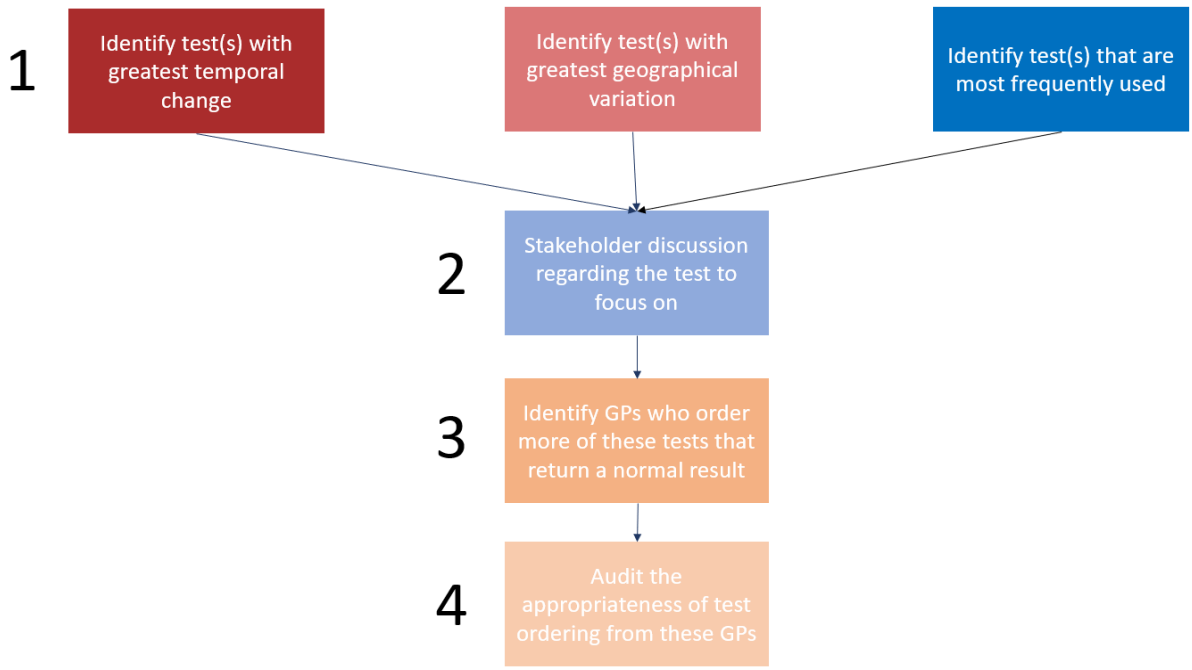


Figure 8.1 A method to identify potentially inappropriate test use

8.5.1.1 Step 1 Identifying tests with the greatest temporal increase, between-practice variation and frequency of use

This step uses numerous indirect markers to identify tests that may be subject to overuse. It ultimately generates a list of tests that may benefit from further investigation. In my thesis, chapters 3 and 4 provided the methods and results to achieve this step. In chapter 3, I identified the tests with the greatest temporal change in use from 2000/1 to 2015/16. For the following discussion, I focus on the period after 2004/5 as there are no concerns about coding inconsistencies and the period after the QOF is more likely to reflect current practice. Table 8.1 shows the top five tests in terms of temporal increase in use from UK primary care from 2004/5 to 2015/16. I also identified the tests that were the most frequently used over the most recent year of the study period (2015/16); table 8.2.

Table 8.1 Top 5 tests with greatest temporal change in use (2004/5 to 2015/16)

Test	**Rank of tests: Temporal change in use (2004/5 to 2015/16)**
Vitamin D	1
Knee MRI	2
Pelvic CT	3
Brain MRI	4
Urine Albumin-Creatinine Ratio	5

Table 8.2 Top 5 most frequently used tests (2015/16)

Test	**Rank of tests: Frequency of use (2015/16)**
Renal Function Tests	1
Full Blood Count	2
Liver Function Tests	3
Lipid tests	4

Thyroid Function tests	5

In Chapter 4 I identified the tests with the greatest between-practice variation in use. Table 8.3 shows the top 5 most variable tests. As described in Chapter 4, tests are ranked according to their coefficient of variation.

Table 8.3 Top 5 tests with most between-practice variation

Test	**Rank of tests: Between-practice variation of use (2015/16)**
Non-illicit drug monitoring	1
Urine Microalbumin	2
Pelvic CT	3
Pap Smear	4
Troponin	5

To identify the test(s) with the greatest combination of a) temporal change in use, b) between-practice variation in use and c) frequency of use, I ranked all 44 tests for these three categories. I then combined each test's rank from these categories and reranked them according to their combined score.

The test with the lowest combined score represents the test with the greatest combination of temporal increase, between-practice variation and frequency of use. From my data (CPRD), I found that vitamin D tests were ranked 1st in terms of temporal change, 21st in terms of frequency of use and 8th in terms of between-practice variation, giving a total rank score of 30 (the lowest of the 44 tests). Table 8.4 shows the 10 tests with the lowest combined score, i.e. the most variable (appendix 8 shows the full list of 44 tests).

Table 8.4 Top 10 tests for combined score

Test	**Rank of tests for the combination of temporal change, frequency of use and between-practice variation in use**
Vitamin D	1
Clotting tests	2
Vitamin B12	3
Bone Profile tests	4
C-Reactive Protein (CRP)	5
Urine Albumin:Creatine ratio	6
Iron	7
Pelvic CT	8
Brain MRI	9
Prostate Specific Antigen	10

Throughout the sections *8.5.1.1 to 8.5.1.4*, I will work through a theoretical example to help explain the four steps of my proposed approach. I will use the example of a UK CCG that wants to identify overtesting.

If the CCG chose to follow my proposed approach, they initially would rank all the tests available to primary care physicians in order of greatest to least a) temporal increase, b) between-practice variation and c) frequency of use. As I have explained in the previous paragraph, they would then determine which of these tests has the lowest combined rank of all three of these variables. They may then highlight the top 10 "most variable" tests to move to the second step. For the sake of discussion, let us imagine that the CCG generated the same top 10 tests as I did (table 8.4).

8.5.1.2 Step 2 Stakeholder involvement

An important limitation of the first step is there may be justifiable reasons why some tests have been subject to a sharp temporal increase in use or marked between-practice variation. The second step allows a stakeholder discussion to explore these issues and ultimately decide on the test they wish to pursue further. Specifically, this step facilitates a discussion about a) local factors that may favour the selection of one test over another and b) reasons that may indicate that between-practice variations and/or temporal increases in use of specific tests may be justified.

Local factors may include the costs of tests (for example, the fifth test may be much more expensive that the first and hence represent greater savings if overuse was mitigated). Local factors may also include varying disease prevalence: a disease may have a lower prevalence in certain areas of the UK and hence mitigating potential overuse of a certain test may be of less relevance to a certain CCG.

At this stage, stakeholders may also discuss the between-practice variation and/or temporal increase in specific tests, for which there may be appropriate reasons. For example, the 2011 World Health Organization guideline on diabetes[25] concluded that HbA1c tests can be used to diagnose, as well as, monitor diabetes. It is likely that this guideline change contributed to the sharp temporal increase in HbA1c test use after 2011. In regard to between-practice variation, varying disease and/or risk factor prevalence may lead to differing rates of test use. For instance, there is marked variation in the use of novel oral anticoagulants (NOACs) around the UK,[26] which may explain the between-practice variation in the use of clotting tests (NOACs, unlike warfarin, do not require routine monitoring tests).

To return to our theoretical worked example, a UK CCG may assess the laboratory costs of all ten tests generated from step 1 (table 8.4) and decide that they want to focus on CRP tests, as they are the most expensive in their laboratory. CRP tests would then become the focus for step 3.

8.5.1.3 Step 3 Identifying General Practices that order a lower proportion of "abnormal" tests

In the first step, a ranking of tests that may be subject to overuse is generated. In the second step, this list is refined. The third step focuses on one test (but users may decide to focus on multiple tests) and identifies general practices that may benefit from further assessment of this test use.

This step identifies general practices that ordered a lower proportion of tests that return an "abnormal" result (i.e. outside the reference range) once differences in practice populations have been accounted for. For instance, in the worked example, the third step would identify general practices that order a significantly lower proportion of CRP tests that return an "abnormal" result compared with other practices. Chapter 5 details how this was performed.

To continue the worked example, a UK CCG would determine, for all the general practices within its jurisdiction, the proportion of all CRP tests ordered that returned an abnormal result. The CCG would then identify general practices that ordered a proportion of abnormal tests 3 standard deviations below the mean of other practices, once differences in practice populations were statistically accounted for.

8.5.1.4 Step 4 Auditing test use against an appropriateness criterion

The first and third steps are *indirect* markers of inappropriateness, that is, in isolation these markers do not reflect inappropriateness; they highlight variation, which may or may not be reasonable. In a sense, these steps identify tests and general practices that would benefit from further, *direct* assessment. The differences between indirect and direct markers of inappropriateness are explained in *Chapter 2 Background* and in *8.4.5 Limitations of the 4-step approach* (below).

Step four focuses on tests identified from the first two steps and the general practices identified in the third step. In the fourth step, the appropriateness of test use from the identified general practices (from step 3) is determined by comparing the individual patient data (IPD) of those who have undergone the test of interest against a national or international guideline. For instance, the UK's National Institute for Health and Care Excellence (NICE) recommends ordering serum iron measurements only when "a full blood count (FBC) and other haematological indices suggest iron deficiency".[27] If, after interrogation of the IPD, a patient had undergone iron studies without a prior FBC and other

haematological indices suggesting iron deficiency, this would be classified as inappropriate overuse. It should be noted that in some clinical circumstances a non-conventional reference range is recommended for certain patient groups. This is true of the European Society of Cardiology (ESC) 2016 guidelines on acute and chronic heart failure (see page 194, section 8.5.2, Limitations of the 4-step approach).

To achieve this last step, the UK CCG, which I have used as a theoretical example throughout these sections, would audit the patient records for patients that had undergone a CRP test, from general practices that were identified from step 3.

8.5.2 Limitations of the 4-step approach

First, it is important to note that my proposed approach requires further work and validation. I have described it in the discussion chapter of my thesis deliberately; it is an idea to move forward in the future.

My four-step approach has not been tested regarding user feasibility nor has it been validated. Before implementation can be considered, users' acceptability of the approach and their perceptions of its feasibility need to be determined. Furthermore, ideally it should be validated: to determine how accurate it is at identifying overuse. I elaborate on these points in *8.4.6 Next Steps*.

With the above in mind, the four steps I propose are not designed to be prescriptive. Even if the above is achieved – the approach is deemed feasible by users and it is validated – I do not envision it to be used rigidly. In fact, I try to reflect this in step 2; it is deliberately open, so users can input their own knowledge and expertise. The identification of waste in a healthcare system is complex, the proposed process is designed to offer users with a structured formula. It may prove with time – and use – that users rely on some aspects more than others or skip certain steps and it is likely that the process will continue to be refined with time.

A limitation of the first step is the equal weighting of temporal increase, between-practice variation, and frequency of use. As described in 8.4.1 Step 1 of the process, I generated a list of tests that had

the lowest combined ranks of these three factors. This is one approach for generating a list of tests that may be subject to overuse; however, there may be other ways to weigh these variables differently. I hope to explore other approaches in the future through qualitative interviews with other researchers (see 8.4.6 Next steps).

Furthermore, it should be noted that steps 1 and 3 are indirect markers of potential waste. As mentioned in section *8.4 Strengths and limitations*, they do not imply inappropriateness in isolation and should be considered 'screening tests' for inappropriateness, requiring further assessment. The fourth step, which considers the indication for the test request and IPD, is required to more definitively determine whether tests are being overused.

A limitation of the third step is the assessment of tests predominately used to monitor disease. Theoretically, a test used to monitor disease should not be ordered except in patients with a confirmed diagnosis. Thus, tests used for monitoring should mostly return a result outside the reference range. Consider the example of urine albumin creatinine ratio (ACR). Urine ACR is predominately a monitoring test used to measure the progression of chronic kidney disease (CKD). The UK's Quality Outcomes Framework (QOF) records "the percentage of patients on the CKD register whose notes have a record of a urine albumin:creatinine ratio test in the preceding 12 months".[28] If urine ACR tests are ordered only in patients with CKD, most of the tests will be outside the reference range and it is therefore plausible that there will be very little variation (of the proportion abnormal) between general practices. If this is the reality, it will be difficult to identify general practices that are ordering a proportion of "abnormal" tests beyond the 99.8% control limits (3 standard deviations). Although this is clearly a limitation of this step, I believe this problem is unlikely for three reasons.

First, there is a paucity of tests that are exclusively used for monitoring. Even tests that are predominately used for monitoring – e.g. urine ACR – are often used in a way that is not congruent with the relevant guidelines,[4] e.g. they are used diagnostically or for screening.

Secondly, tests that are predominately monitoring tests are often recommended for screening a patient with a known risk factor. Urine ACR is a good example of this. The QOF records both "the

percentage of patients with a new diagnosis of hypertension who have a record of urinary albumin: creatinine ratio test in the three months before or after the date of entry to the hypertension register"[29] and "the percentage of patients with diabetes who have a record of an albumin:creatinine ratio (ACR) test in the preceding 15 months".[30] Thus, a urine ACR test is recommended for screening for a disease (CKD) in patients with known risk factors (hypertension and diabetes) and this should introduce variation in the proportion of tests that return an "abnormal" result between practices.

Lastly, chronic disease does not universally worsen. In some cases, for instance type 2 diabetes, the disease improves. In these cases, a monitoring test will start returning "normal" results and lead to more variability in the proportion of abnormal monitoring tests.

Furthermore, even if there is minimal variation between general practices in the proportion of monitoring tests that return abnormal results, the third step of my approach should still assist in identifying practices that would benefit from further assessment, such as those that order more monitoring tests than their colleagues (x-axis on figure 5.2 (chapter 5)).

It should also be acknowledged that the dichotomisation of test results into "normal" or "abnormal" may prove to be too simplistic. It is reasonable to expect some tests that return a result outside the reference range may, in the context of an individual patient, be interpreted by a GP as normal, or at least unchanged. This is probably most relevant for laboratory monitoring tests. For instance, a patient with chronic heart failure may return a B-type natriuretic peptide (BNP) test result that is outside the reference range but below their most recent reading. This issue is similar to the limitations previously addressed regarding chronic disease monitoring (page 192). Similarly, there may be certain patients where the reference ranges for common tests are adjusted. The best example of this is iron tests in patients with heart failure. The European Society of Cardiology (ESC) recommends treating heart failure patients with iron supplementation at levels considered normal in non-heart failure patients. The recommended thresholds they suggest are a "serum ferritin <100 μg/L, or ferritin between 100 – 299 μg/L and transferritin saturation <20%".[43] This could prove problematic for the third step of my approach because there may be tests that by conventional reference ranges are "normal", but given the

specific clinical circumstances should actually be interpreted and actioned as if they are abnormal. Although this is clearly a limitation, there are numerous reasons why this is unlikely to jeopardise the interrogatory of my approach. 1. There are a paucity of clinical situations where adjusted reference ranges are currently recommended and 2. Step four is an audit of test use against individual patient data (IPD). This step should address the above concerns; in the audit process of this step, the use of IPD will allow context-dependent assessment of the appropriateness of test use.

A further limitation is the quality of primary care electronic records to adequately quantify overuse. This limitation is relevant to the fourth step of my proposed approach. I have outlined the limitations above in section *8.4 Strengths and Limitations*.

As my thesis broadly concerns NHS sustainability, the omission of any data relating to test costs is clearly a limitation. While I tried to address this in step 2 of my four-step approach, a more thorough explanation is required. At the protocol stage of the research that supports chapters 3 and 4 (CPRD projects), there were many discussions surrounding the feasibility of including data surrounding costs. Ultimately, it was not possible to include a formal assessment of test costs for numerous reasons. First, the CPRD does not include data relating to test cost. Second, as our analyses assessed trends over time (chapter 3) and variation between practices (chapter 4), I would have required test cost data for 16 different years in time and for over 700 general practices across the country. No data source to address these geospatial concerns emerged and to my knowledge, there remains no apparent way to address this. Third, the CPRD anonymises it's data, thus it was not possible for me to identify the particular location of each general practice, nor was I able to identify the CCG a practice was within. In future work I hope to develop a feasible way to include the cost of tests into the four-step approach.

As I have mentioned throughout this chapter, I envision that my proposed approach will undergo iterations with time. Importantly, throughout my thesis this has already occurred. In the early phases, I aimed to identify both overuse and underuse of tests. The first approach I designed was equipped to identify both inappropriate overuse and underuse of tests. My decision to focus on overuse was led by the results of my studies. I was particularly influenced by the results described in Chapter 3: of the 44

tests I assessed, the use of only one fell significantly from 2000/1 to 2015/16, similarly only four tests fell significantly from 2004/05 to 2015/16. I initially proposed that a decrease in test use over time, without clear explanation, may be an indirect marker of potential underuse. However, given my results suggested that most tests increased over time, overuse seemed to be much more prevalent than underuse, or at least the marker I was using to detect underuse and overuse (temporal change in use) was more sensitive at detecting overuse. An additional reason for focusing on overuse was my desire to detect substantial amounts of wasteful use. This desire led me to incorporate the box "identify test(s) that are most frequently used" into step 1 (figure 8.1). I included this because I wanted to identify tests that were being frequently used inappropriately. Reduced use of tests that were being frequently overused would, I concluded, lead to larger savings for the NHS. The initial approach I proposed (without frequency of test use), may still be of value to policy makers, but currently the results from my studies led me to believe that identifying overuse is more appropriate.

8.6 Implications for future research

A major limitation of my 4-step approach is that guideline recommendations for tests are supported by poor-quality evidence. This jeopardises step 4 of my approach: direct assessment of appropriateness. Without high-quality evidence that a test is accurate or improves a patient's outcomes, it is difficult to confidently classify the use of that test as truly inappropriate. Consider two contrasting examples: the use of imaging for patients with non-red flag low back pain and a chest x-ray to confirm or refute the diagnosis of heart failure. A 2009 systematic review and meta-analysis examined the effect of imaging on quality of life and other patient outcomes in patients with non-red flag low back pain.[31] This review showed no difference in patient-important outcomes between those who underwent imaging and those who didn't (both at short-term and long-term follow up). The quality of this research and its focus on patient-important outcomes allows us to confidently state that the use of imaging in non-red flag low back pain is inappropriate. In contrast, the usefulness of a chest x-ray (CXR) in diagnosing heart failure (HF) is much less clear. Currently the European Society of Cardiology (ESC) states that "a chest x-ray is of limited use in the diagnostic work-up of patients with

suspected HF."[32] However, they go on to state that a CXR "may, however, show pulmonary venous congestion or oedema in a patient with HF, and is more helpful in the acute setting than in the non-acute setting".[32] The evidence the ESC used to support this recommendation was a narrative review[33] and a primary study that assessed different characteristics (including "flow cephalization on chest x-ray") that are associated with impaired systolic left ventricular function in patients with heart failure.[34] It is therefore difficult to confidently state, from either the guideline recommendation or its underlying evidence, whether a chest x-ray in patients with suspected heart failure is appropriate or inappropriate. As we detailed in Chapter 7, this is a common scenario for many diagnostic tests. To be able to complete step 4 of our process the evidence base supporting guideline recommendations for tests needs to improve.

The most effective way to improve the evidence base for tests is by performing randomised controlled trials (RCTs), in which the use of a test is randomised, assessed as an intervention, and patient-important outcomes are measured. This was the design used in the RCTs that assessed imaging for patients with non-red flag low back pain. This view, of assessing tests in RCTs, is supported by many organisations including NICE, The American College of Physicians, The BMJ, and the WHO.[35–38]

However, even in the absence of RCTs, guideline recommendations for tests can be supported by high-quality evidence: they can be based on diagnostic accuracy studies. In these studies, one group of patients undergoes an index test (the test in which accuracy is to be determined) and a reference standard test (ideally a gold standard test; that is, the test that is best at classifying patients as diseased or not). These studies produce measures of diagnostic accuracy, most commonly sensitivity and specificity, of the index test. Such studies, if conducted with minimal bias, can show that a test is very useful at ruling a disease in or out, and can thus support a high-quality guideline recommendation.

Unfortunately, RCTs evaluating tests are rare, as are high-quality diagnostic accuracy studies (as I have shown in Chapter 7). There is a need to improve the evidence base for tests. The absence of high-quality evidence also makes the identification and mitigation of overtesting difficult; in the

absence of high-quality evidence, it will remain difficult to define when test use is, and when it is not, appropriate.[3,22]

Furthermore, there are specific future research implications from chapter 5 (*Variation in diagnostic test requests and outcomes)*. It would be advantageous to compare the test use of general practices greater than 3 standard deviations (SDs) from the mean with general practices within 2 SDs of the mean. Specifically, it would be of value to determine the proportion of test use that was ordered incongruent with NICE guidelines for both of these groups (practices within 2 SDs and practices outside 3 SDs).

My research has other implications for future research. I am yet to gather external input on my proposed 4-step approach to identify potential overtesting; it would be of value to gather the thoughts of policy makers, general practitioners, patients and other academics. To achieve this, I hope, post-doctorally, to conduct qualitative interviews with a) policy makers, b) general practitioners (GPs), c) patients and d) other academics in the future.

Specifically, I would be interested to address the following questions (among others):

- Policy makers:
    - How useful would the 4-step approach be to them?
    - Would they use it?
    - If they did use it, would it be feasible and easy for them to use?
    - Could they suggest improvements?
- General Practitioners
    - Could my approach, and the data contained in it, be of value to GPs?
    - Would a GP be interested in a) which tests they order most and b) which tests have had the greatest increase in use (compared with GPs in other practices, after adjustment for differences in practice populations)?

    - Would a GP be concerned if they were ordering a much lower proportion of tests that returned an abnormal result compared with GPs in other practices (after adjustment for practice populations)?
- Patients
    - What do patients make of substantial increases in test use over time?
    - What do patients make of the variation between general practices in both the number of tests ordered and the proportion that returns a normal result?
- Academics
    - Are there any other indirect measures that could be of value?
    - Are there ways of weighting temporal increase, between-practice variation, and frequency of use?
    - Are there any other ways to directly measure appropriateness of the use of tests, aside from guidelines?
    - What do they think is the best way to validate these steps?

Additionally, I am part of the team that is starting to develop OpenPathology.net and OpenImaging.net. Both sites are based on Openprescribing.net: an openly accessible, open source, and open data tool that provides easy access to various analytic approaches to identify outliers for prescribing behaviour in primary care. I, together with Ben Goldacre and his team, am developing OpenPathology and OpenImaging; these data tools will provide open, online, live and interactive data tools to assess test use in UK primary care. We are using methods I developed in this DPhil; for instance, we hope to provide a live interface to explore temporal trends in test use, between practice and regional variation and variation between practice in the proportion of tests that return an abnormal result. We hope these tools will be used by clinicians, patients, academics and policy makers. Importantly, these tools will be updated automatically; currently a limitation of my proposed 4-step approach is the amount of work required to achieve each step. OpenPathology and OpenImaging would present the results required for Step 1 and 3 in an automated fashion.

8.5.1 Validation of my approach

After my approach has been refined via the above steps, it will still require validation. The Oxford English Dictionary (OED) defines "to validate" as "to examine for incorrectness or bias; to confirm or check the correctness of."[39] The difficulty in validating my approach is that there is currently no strategy, approach or method to identify overtesting; i.e. there is no "gold standard" approach. This is problematic because I cannot compare my approach to a recognised approach. Nevertheless, below I present a method that I believe represents the best possible validation method for my approach (in its current form) in the absence of a gold standard method.

Like all healthcare interventions, tests are ultimately only useful if they contribute to improved patient outcomes. It is this reasoning that supports the notion that tests should ideally be evaluated in RCTs.[38] Similarly, my 4-step approach is only useful if it helps contribute to a reduction in ineffective use of tests without jeopardising patient outcomes. In a sense, I want to identify the tests that do not positively contribute to patient care. I therefore propose to examine my approach in a cluster RCT (after the expected and detailed iterations occur from further research and work). In this trial, I would implement my 4-step approach as an audit and feedback tool: some Clinical Commissioning Groups (CCGs) (and the General Practices within these CCGs) would be randomised to use my approach and other CCGs would be randomised to usual care. CCGs randomised to use my approach would follow its steps: they would identify a test to focus on, identify general practices that maybe overtesting, and then quantify overtesting against NICE guidelines. Each of the general practices would then be fed back their results: how much they request this test compared to their colleagues, their rate of test use that returns an abnormal result compared to their colleagues and the amount of testing that was not in line with NICE guidelines.

After the feedback has been given to general practices, the following outcomes would be measured prospectively: volume of test use, appropriateness (against NICE guidelines), hospitalisation rates, referral rates, GP consultation rates, and mortality. For the last four outcomes the following patient

groups should be assessed a) all patients registered in a general practice and b) patients in whom the test is indicated or contraindicated. Figure 8.3 displays the proposed method.

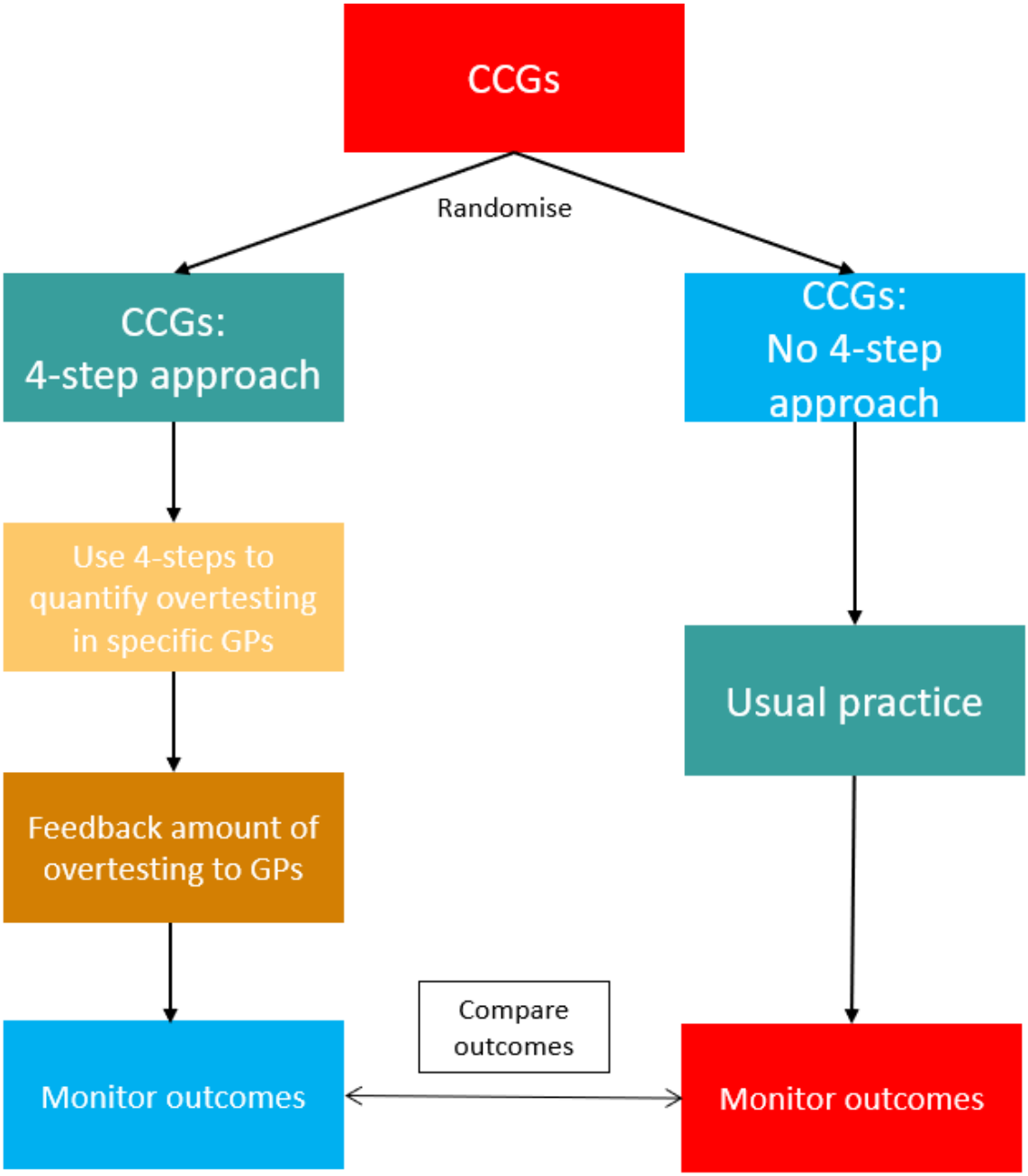


Figure 8.3 Study design for cluster RCT. CCGs: Clinical Commissioning Groups, GPs: General Practices

The assessment of my process in a cluster RCT would assess my 4-step approach as a feedback tool and ascertain whether a reduction in specific test use occurred and if it did occur, did it occur with or without a reduction in patient outcomes (e.g. admission rate, reconsultation rate and mortality). Both OpenPathology.net and OpenImaging.net should assist in the measurement of some of the outcomes from this proposed RCT.

A cluster RCT is one potential way to try and validate my four-step approach. It is likely that with more work post-doctorally other methods will become apparent. Furthermore, more detailed work is required to determine whether a cluster RCT would be feasible or practical. At this stage, my 4-step approach is still an idea I am hoping to move forward post-doctorally.

8.6 My reflections

8.6.1 My research reflections

I arrived in Oxford as a starry-eyed junior doctor. I had some research experience, but I had no experience in what I consider the "research journey". My research experience prior to my DPhil involved being told to execute something, for instance, title and abstract screen these 2000 papers for studies that matches our stated eligibility criteria. I was not capable of independent research. During this DPhil, I have learnt to generate research ideas, design a study to answer a question, acquire the appropriate data, analyse the data and present the results in a coherent way – the "research journey".

I have learnt a tremendous amount. Although I still very much consider my skills as a work in progress, my ability to manage data and run statistical analyses (both through coding in R) has exceed any expectation I had for myself. As has my ability to generate feasible and impactful research ideas and contribute critically and usefully to others' research. I feel equipped with the skills to conduct high quality research.

8.6.2 My personal reflections

In November 2017, Carl Heneghan said to me, having supervised me for over two years, “Jack, has anyone commented about how much you’ve changed? You still get excited as Jack O’Sullivan gets excited, but you are much more measured now.” Ironically, the answer to Carl’s question is no. No one else has commented that I had matured. Although I am not sure anyone is better placed to notice than Carl. On reflection, I recognise what Carl is eluding to. I feel more focused, better able to understand my place in the world and contribute positively to the activities I am pursuing. Undoubtedly, I have lots of growth ahead of me, but it is nice to acknowledge the personal growth that has been achieved during this DPhil.

8.7 Conclusion

Healthcare systems around the world are struggling to remain financial sustainable. Primary care, given its volume of activity, presents the ideal setting in which to mitigate wasteful healthcare resource use. There is very little evidence exploring primary care test use, despite its frequency of use and contribution to expenditure. In this thesis, I have used a variety of biostatistical and meta-research approaches to analyse a variety of data sources to ultimately produce an approach that may help identify potentially inappropriate overuse of tests.

There is still more work required before this process can be implemented, however, each of my results chapters (chapters 3-7) have tested several of the steps of the proposed approach and thus shown, even in its current form, it can produce meaningful results.

References

1 Alderwick H, Robertson R, Appleby J, Dunn P, Maguire D. Better value in the NHS The role of changes in clinical practice. 2015.

2 Gainsbury S. Nuffield Trust: Feeling the crunch: NHS finances to 2020. 2016 https://www.nuffieldtrust.org.uk/research/feeling-the-crunch-nhs-finances-to-2020.

3 Brownlee S, Chalkidou K, Doust J, *et al.* Evidence for overuse of medical services around the world. *Lancet* 2017; **6736**: 1–13.

4 O'Sullivan JW, Albasri A, Nicholson B, *et al.* Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open* 2018; **8**: e018557.

5 Zhi M, Ding EL, Theisen-Toupal J, *et al.* The Landscape of Inappropriate Laboratory Testing: A 15-Year Meta-Analysis. *PLoS One* 2013; **8**: e78962.

6 Gontijo RV, Proietti FA, Santos Amaral CF, De Rezende NA. Appropriateness use of coronary angiography in patients with suspected ischemic heart disease in Brazil. *Int J Cardiol* 2005; **104**: 348–9.

7 Carpeggiani C, Marraccini P, Morales MA, Prediletto R, Landi P, Picano E. Inappropriateness of cardiovascular radiological imaging testing; a tertiary care referral center study. *PLoS One* 2013; **8**: 1–5.

8 Busby J, Schroeder K, Woltersdorf W, *et al.* Temporal growth and geographic variation in the use of laboratory tests by NHS general practices: Using routine data to identify research priorities. *Br J Gen Pract* 2013; **63**: 256–66.

9 Appleby J, Raleigh V, Frosini F, Bevan G, Gao H, Lyscom T. Variations in health care: The good, the bad and the inexplicable. 2011.

10 Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med* 2005; **24**: 1185–202.

11 Spiegelhalter DJ. Mortality and volume of cases in paediatric cardiac surgery: retrospective study based on routinely collected data. *BMJ* 2002; **324**: 261–3.

12 Spiegelhalter DJ. Handling over-dispersion of performance indicators. *Qual Saf Heal Care* 2005; **14**: 347–51.

13 Aylin P, Best N, Bottle A, Marshall C. Following Shipman: A pilot system for monitoring mortality rates in primary care. *Lancet* 2003; **362**: 485–91.

14 Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. *JAMA* 1999; **281**: 1900–5.

15 Grilli R, Magrini N, Penna A, *et al.* Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet (London, England)* 2000; **355**: 103–6.

16 Lenzer J. Why we can't trust clinical guidelines. *BMJ* 2013; **346**.

17 Fitch K, Bernstein SJJ, Aguilar MDD, *et al.* The RAND / UCLA Appropriateness Method User 's Manual. 2001 http://www.rand.org.

18 Mafi JN, McCarthy EP, Davis RB, Landon BE. Worsening trends in the management and treatment of back pain. *JAMA Intern Med* 2013; **173**: 1573–81.

19 Mafi JN, Edwards ST, Pedersen NP, Davis RB, McCarthy EP, Landon BE. Trends in the Ambulatory Management of Headache: Analysis of NAMCS and NHAMCS Data 1999-2010. *J Gen Intern Med* 2015; **30**: 548–55.

20 Elwyn G, Owen D, Roberts L, *et al.* Influencing referral practice using feedback of adherence to NICE guidelines: a quality improvement report for dyspepsia. *Qual Saf Health Care* 2007; **16**: 67–70.

21 Gurzun M-M, Ionescu A. Appropriateness of use criteria for transthoracic echocardiography: are they relevant outside the USA? *Eur Hear J - Cardiovasc Imaging* 2014; **15**: 450–5.

22 Chalmers K, Pearson S, Elshaug AG. Quantifying low-value care: a patient- versus service-centric lens. *BMJ Qual Saf* 2017; **26**: 855–8.

23 Glasziou P, Straus SE, Brownlee S, *et al.* Evidence for underuse of effective medical services around the world. *Lancet* 2017; **390**: 169–77.

24 Saini V, Brownlee S, Elshaug AG, Glasziou P, Heath I. Addressing overuse and underuse around the world. *Lancet* 2017; **6736**: 10–3.

25 World Health Organization. Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. 2011 DOI:WHO/NMH/CHP/CPM/11.1.

26 Millett D. NOAC prescribing varies 16-fold between CCG areas. GP Online. 2016. https://www.gponline.com/noac-prescribing-varies-16-fold-ccg-areas/cv-thromboembolic-disorders/atrial-fibrillation/article/1394082.

27 NICE. Iron tests. Do Not Do Recomm. 2018. https://www.nice.org.uk/donotdo/tests-for-serum-ferritin-in-adults-should-not-be-carried-out-unless-a-full-blood-count-and-other-haematological-indices-suggest-iron-deficiency.

28 National Institute for Health and Care Excellence. QOF Indicator NM109 for Chronic Kidney Disease. 2015.

29 National Institute for Health and Care Excellence. QOF Indicator NM75 for Hypertension. 2014.

30 National Institute for Health and Care Excellence. QOF Indicator NM59 for Diabetes. 2016.

31 Chou R, Fu R, Carrino JA, Deyo RA. Imaging strategies for low-back pain: systematic review and meta-analysis. *Lancet* 2009; **373**: 463–72.

32 Ponikowski P, Voors AA, Anker SD, *et al.* 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J* 2016; **37**: 2129–2200m.

33 Hawkins NM, Petrie MC, Jhund PS, Chalmers GW, Dunn FG, McMurray JJV. Heart failure

and chronic obstructive pulmonary disease: Diagnostic pitfalls and epidemiology. *Eur J Heart Fail* 2009; **11**: 130–9.

34 Thomas JT, Kelly RF, Thomas SJ, *et al.* Utility of history, physical examination, electrocardiogram, and chest radiograph for differentiating normal from decreased systolic function in patients with heart failure. *Am J Med* 2002; **112**: 437–45.

35 Siontis KC, Siontis GCM, Contopoulos-Ioannidis DG, Ioannidis JPA. Diagnostic tests often fail to lead to changes in patient outcomes. *J Clin Epidemiol* 2014; **67**: 612–21.

36 Ioannidis JPA. Appropriate vs Clinically Useful Diagnostic Tests. *Jama* 2013; **173**: 2011–3.

37 Lijmer JG, Bossuyt PMM. Various randomized designs can be used to evaluate medical tests. *J Clin Epidemiol* 2009; **62**: 364–73.

38 Schünemann HJ, Oxman AD, Brozek J, *et al.* GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008; **336**: 0–b.

39 Oxford English Dictionary. 'validate, v.' 2018. http://www.oed.com/view/Entry/221192?rskey=DfCePo&amp.

40 Elshaug AG, Watt AM, Mundy L, Willis CD. Over 150 potentially low-value health care practices: an Australian study. *Med J Aust* 2012; **197**: 556–60.

41 Goldzweig CL, Orshansky G, Paige NM, *et al*. Electronic health record-based interventions for improving appropriate diagnostic imaging: a systematic review and meta-analysis. *Ann Intern Med.* 2015; 162(8): 557-65.

42 Mafi JN, Parchman M. Low-value care: an intractable global problem with no quick fix. *BMJ Qual Saf* 2018; 27(5):333-336.

43 Ponikowski P, Voors AA, Anker SD, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J* 2016;37:2129–2200m. doi:10.1093/eurheartj/ehw128

Appendix Chapter 1

1.1 DPhil publications

1. **O'Sullivan JW**, Heneghan C, Perera R, Oke J, Aronson JK, Shine B, Goldacre B. Variation in diagnostic test requests and outcomes: a preliminary metric for OpenPathology.net. Sci Rep 2018;8:4752. doi:10.1038/s41598-018-23263-z.
2. **O'Sullivan JW**, Albasri A, Nicholson B, Perera R, Aronson J, Roberts N, et al. Overtesting and undertesting: a systematic review and meta-analysis. BMJ Open 2018;8:e018557. doi:10.1136/bmjopen-2017-018557.
3. **O'Sullivan JW**, Albasri A, Koshiaris C, et al. Diagnostic test guidelines based on high-quality evidence had greater rates of adherence : a meta-epidemiological study. J Clin Epidemiol 2018;103:40–50. doi:10.1016/j.jclinepi.2018.06.013
4. **O'Sullivan JW**, Stevens S, Hobbs FDR, Salisbury C, Little P, Goldacre B, Bankhead C, Aronson JK, Perera R, Heneghan C. Temporal trends in the use of tests in UK primary care: a retrospective analysis of 250 million tests, 2000 – 2015. BMJ (Accepted)
5. **O'Sullivan JW**, Stevens S, Hobbs FDR, Salisbury C, Little P, Goldacre B, Bankhead C, Aronson JK, Heneghan C, Perera R. Identifying tests with the greatest geographical variation in their use. BMC Medicine (Provisionally accepted).

1.2 Non-DPhil Publications

6. **O'Sullivan JW,** Muntinga T, Grigg S, Ioannidis JPA. Prevalence and Outcomes of Incidental imaging findings: An Umbrella Review. BMJ 2018; 361: k2387.
7. **O'Sullivan JW**, Banerjee A, Heneghan C, Pluddemann A. Verification bias. BMJ Evidence-Based Med 2018;23:54 LP-55.
8. **O'Sullivan JW.** Developing guideline recommendations for diagnostic tests. 2018. Chapter in WHO Handbook for Guideline Development.

9. **O'Sullivan JW**, Harvey RT, Glasziou PP, McCullough A. Written information for patients (or parents of child patients) to reduce the use of antibiotics for acute upper respiratory tract infections. Cochrane Database Syst Rev. 2016 Nov 25;11:CD011360. Review. PMID: 27886368
10. **O'Sullivan J**. Controversies in PSA screening. Evid Based Med. 2017 Dec;22(6):198. doi: 10.1136/ebmed-2017-110858. Epub 2017 Nov 9. PMID: 29122825
11. **O'Sullivan JW,** Del Mar C, Barnes A, Marron L. Acupuncture in Australia. Focus on Alternative and Complementary Therapies. 2016 Feb; 21(1):22-24. doi: 10.1111/fct.12228
12. Albasri A, **O'Sullivan JW**, Roberts NW, Prinjha S, McManus RJ, Sheppard JP. J Hypertens. A comparison of blood pressure in community pharmacies with ambulatory, home and general practitioner office readings: systematic review and meta-analysis. 2017 Oct;35(10):1919-1928. doi: 10.1097/HJH.0000000000001443. PMID: 28594707
13. Warriner DR, **O'Sullivan JW**. Has too much cardiology been sent into the appropriateness ORBITA? BMJ Evidence-Based Med 2018;23:48 LP-49.
14. Collins DR, Albasri A, **O'Sullivan J**, Bobrovitz N. Blood pressure lowering for cardiovascular disease. Lancet. 2016 Jul 9;388(10040):125. doi: 10.1016/S0140-6736(16)30971-0. PMID: 27411869
15. Aldous S, Mark Richards A, George PM, Cullen L, Parsonage WA, Flaws D, Florkowski CM, Troughton RW, **O'Sullivan JW**, Reid CM, Bannister L, Than M. Comparison of new point-of-care troponin assay with high sensitivity troponin in diagnosing myocardial infarction. Int J Cardiol. 2014 Nov 15;177(1):182-6. doi: 10.1016/j.ijcard.2014.09.026. Epub 2014 Sep 28. PMID: 25499373
16. Onakpoya I, **O'Sullivan J**, Heneghan C, Thompson M. The effect of grapefruits (Citrus paradisi) on body weight and cardiovascular risk factors: A systematic review and meta-analysis of randomized clinical trials. Crit Rev Food Sci Nutr. 2017 Feb 11;57(3):602-612. Review. PMID: 25880021
17. Onakpoya IJ, **O'Sullivan J**, Heneghan CJ. The effect of cactus pear (Opuntia ficus-indica) on body weight and cardiovascular risk factors: a systematic review and meta-analysis of randomized

clinical trials. Nutrition. 2015 May;31(5):640-6. doi: 10.1016/j.nut.2014.11.015. Epub 2014 Dec 19. Review. PMID: 25837206

18. Onakpoya IJ, **O'Sullivan J**, Thompson MJ, Heneghan CJ. The effect of wind turbine noise on sleep and quality of life: A systematic review and meta-analysis of observational studies. Environ Int. 2015 Sep;82:1-9. doi: 10.1016/j.envint.2015.04.014. Epub 2015 May 16. Review. PMID: 25982992
19. Nunan D, **O'Sullivan JW**, Heneghan C, Pluddemann A, Aronson J, Mahtani K. Ten essential papers for the practice of evidence-based medicine. Evid Based Med. 2017 Dec;22(6):202-204. doi: 10.1136/ebmed-2017-110854. Epub 2017 Nov 23. PMID: 29170157
20. Oke J, **O'Sullivan JW**, Nicholson B, Perera R. Temporal trends in incidence and mortality of common cancers in the UK. *Scientific Reports*.
21. Nicholson B, Hamilton W, **O'Sullivan JW**, Aveyard P, Hobbs FDR. Weight loss as a predictor of cancer: a systematic review and meta-analysis. Br J Gen Pract.
22. Brodersen J, Schwartz LM, Heneghan C, **O'Sullivan JW**, Aronson J, Woloshin S. Overdiagnosis: what it is and what it isn't. BMJ EBM 2018 Jan; 23: 1–3. doi: 10.1136/ebmed-2017-110886
23. Heneghan C, Spencer EA, Bobrovitz N, Collins DR, Nunan D, Plüddemann A, Gbinigie OA, Onakpoya I, **O'Sullivan J**, Rollinson A, Tompson A, Goldacre B, Mahtani KR. Lack of evidence for interventions offered in UK fertility centres. BMJ. 2016 Nov 27;355:i6295. doi: 10.1136/bmj.i6295. PMID: 27890864
24. Coombs C, Onakpoya I, Mahtani K, Aronson J, **O'Sullivan J**, Pluddemann A, Heneghan C. What evidence affects clinical practice? An analysis of Evidence-Based Medicine commentaries. Evid Based Med. 2017 Dec;22(6):197. doi: 10.1136/ebmed-2017-110856. Epub 2017 Nov 17. PMID: 29150553

1.3 NIHR Grant Award Letter

NHS
National Institute for Health Research

School for Primary Care Research
Nuffield Department of Primary Care Health Sciences
Oxford University
Radcliffe Observatory Quarter
Woodstock Road, Oxford
OX2 6GG

Tel: 01865 617188
Web: www.spcr.nihr.ox.ac.uk

Dr Jack O'Sullivan
University of Oxford
Nuffield Department of Primary Care Health Sciences
Radcliffe Observatory Quarter
Woodstock Road
Oxford OX2 6GG

11 July 2017

Dear Dr Sullivan

RE: Partner Application to the National Institute for Health Research School for Primary Care Research (NHIR SPCR)
Grant Reference Number: 386
Quantifying changes and appropriateness of UK primary care clinical activity and workload: A 15-year retrospective analysis of diagnostic test use.

The National Institute for Health Research ("NIHR") has, agreed to provide funds, reference :SPCR-2014-10043 (the "Grant") for the National Institute for Health Research School for Primary Care Research ("NIHR SPCR") to be used to increase the evidence base for primary care practice in England by commissioning and conducting world class research. The Grant is to be administered by Oxford under the terms of an agreement between Oxford and Secretary of State for Health (the "Authority") dated with effect from October 2015 (the "Main contract").

I am delighted to inform you that your recent Project Application to SPCR Board for funding for the Research Project named, has been successful. The NIHR SPCR will provide funding of no more than £89,203 over 12 months for the project described in your Project Application.

By signing this Agreement, you agree to carry out the Research Project in accordance with:

(a) the work plan submitted as part of the Project Application, in the form in which it was approved by the Senior Management Team; and

(b) the terms and conditions of a Partnership Agreement to which your institution is a party dated with effect from October 1, 2015; and

(c) the following additional conditions specified by the Senior Management Team:

Appendix Chapter 3

3.1 ISAC protocol

Below is the ISAC protocol (ISAC protocol number 17_06R). This protocol supports the work in Chapters 3 and 4 of my thesis. Please note that some of the outcomes listed in this protocol concerned the wider project and were not relevant to this thesis (and therefore were not included in this thesis).

SECTION A: GENERAL INFORMATION ABOUT THE PROPOSED RESEARCH STUDY
1. Study Title[§] (*Please state the study title below)* Quantifying variation in UK primary care test use: A 15 year retrospective CPRD analysis of temporal and geographical variation in test use. [§]*Please note: This information will be published on the CPRD's website as part of its transparency policy.*
2. Has any part of this research proposal or a related proposal been previously submitted to ISAC? Yes [*] ☐ No ☒ *****If yes, please provide the previous protocol number/s below. Please also state in your current submission how this/these are related or relevant to this study.*
3. Has this protocol been peer reviewed by another Committee? (e.g. grant award or ethics committee) Yes[*] ☒ No ☐

If Yes, please state the name of the reviewing Committee(s) below and provide an outline of the review process and outcome as an Appendix to this protocol :

A version of this protocol was submitted to the Internal funding round 13 of the School for Primary Care Research (Department of Primary Care Health Sciences, University of Oxford). That protocol was not prioritised for funding. A version of this protocol was submitted to the External funding round 14 of the School for Primary Care Research and was prioritised for Round 2 of the Funding Round.

4. Type of Study (please tick all the relevant boxes which apply)

Adverse Drug Reaction/Drug Safety	☐	Drug Effectiveness	☐
Drug Utilisation	☐	Pharmacoeconomics	☐
Disease Epidemiology	☐	Post-authorisation Safety	☐
Health care resource utilisation	☐	Methodological Research	☐
Health/Public Health Services Research	☒	Other*	☐

**If Other, please specify the type of study in the lay summary*

5. Health Outcomes to be Measured[§]

[§]*Please note: This information will be published on CPRD's website as part of its transparency policy.*

Please summarise below the primary/secondary health outcomes to be measured in this research protocol:

Objective 1: To determine the temporal change in test use in UK primary care from 2000 – 2015

Aim 1.1: To examine changes in total test ordering in UK primary care from 2000-2015

Aim 1.2: To examine changes in use of 25 specific tests in UK primary care from 2000 – 2015 (see Appendix A for lists of tests)

Objective 2: Objective 2: To determine geographical differences in test use in UK primary care

Aim 2.1: To determine if geographical variation in total test use exists across the UK

Aim 2.2: To determine if geographical variation in use of 25 specific tests exists across the UK

6. Publication: This study is intended for (please tick all the relevant boxes which apply)**:**

Publication in peer-reviewed journals	☒	Presentation at scientific conference	☒
Presentation at company/institutional meetings	☐	Regulatory purposes	☐
Other*	☐		

**If Other, please provide further information:*

SECTION B: INFORMATION ON INVESTIGATORS AND COLLABORATORS

7. Chief Investigator[§]

Professor Rafael Perera

Professor of Medical Statistics, Nuffield Department of Primary Care Health Sciences, University of Oxford.

rafael.perera@phc.ox.ac.uk

Please state the full name, job title, organisation name & e-mail address for correspondence - see guidance notes for eligibility. Please note that there can only be one Chief Investigator per protocol.

[§]*Please note: The name and organisation of the Chief Investigator and will be published on CPRD's website as part of its transparency policy*

CV has been previously submitted to ISAC ☒ **CV number:** 426_16CESL

A new CV is being submitted with this protocol ☐

An updated CV is being submitted with this protocol ☐

8. Affiliation of Chief Investigator (full address)

Department of Primary Care Health Sciences, University of Oxford.

Radcliffe Primary Care Building

Radcliffe Observatory Quarter

Woodstock Rd, Oxford

OX2 6GG

9. Corresponding Applicant[§]

Please state the full name, affiliation(s) and e-mail address below:

Dr Jack O'Sullivan

Nuffield Department of Primary Care Health Sciences, University of Oxford

jack.osullivan@phc.ox.ac.uk

[§]*Please note: The name and organisation of the corresponding applicant and their organisation name will be published on CPRD's website as part of its transparency policy*

Same as chief investigator ☐

CV has been previously submitted to ISAC ☒ **CV number:** 214_17

A new CV is being submitted with this protocol ☐

An updated CV is being submitted with this protocol ☐

10. List of all investigators/collaborators[§]

Please list the full name, affiliation(s) and e-mail address* of all collaborators, other than the Chief Investigator below:

[§]*Please note: The name of all investigators and their organisations/institutions will be published on CPRD's website as part of its transparency policy*

Title: Dr

Surname: O'Sullivan

Forename: Jack

University: University of Oxford

Email: jack.osullivan@phc.ox.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 214_17

A new CV is being submitted with this protocol ☐

An updated CV is being submitted with this protocol ☐

Title: Professor

Surname: Heneghan

Forename: Carl

University: University of Oxford

Email: carl.heneghan@phc.ox.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 213_17

A new CV is being submitted with this protocol ☐

An updated CV is being submitted with this protocol ☐

Title: Dr

Surname: Bankhead

Forename: Clare

University: University of Oxford

Email: clare.bankhead@phc.ox.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 396_15CESL

A new CV is being submitted with this protocol ☐

An updated CV is being submitted with this protocol ☐

Title: Professor

Surname: Perera

Forename: Rafael

University: University of Oxford

Email: rafael.perera@phc.ox.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 426_16CESL

A new CV is being submitted with this protocol ☐

An updated CV is being submitted with this protocol ☐

Title: Professor

Surname: Little

Forename: Paul

University: University of Southampton

Email: P.Little@soton.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 693_16

A new CV is being submitted with this protocol ☐

An updated CV is being submitted with this protocol ☐

Title: Professor

Surname: Salisbury

Forename: Chris

University: University of Bristol

Email: C.Salisbury@bristol.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 011_17

A new CV is being submitted with this protocol ☐

An updated CV is being submitted with this protocol ☐

Title: Miss

Surname: Stevens

Forename: Sarah

University: University of Oxford

Email: sarah.stevens@phc.ox.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 042_15ESL

A new CV is being submitted with this protocol ☐

An updated CV is being submitted with this protocol ☐

Title: Dr

Surname: Mukhtar

Forename: Toqir

University: University of Oxford

Email: toqir.mukhtar@phc.ox.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 212_17

A new CV is being submitted with this protocol ☐

An updated CV is being submitted with this protocol ☐

Title: Professor

Surname: Hobbs

Forename: Richard

University: University of Oxford

Email: richard.hobbs@phc.ox.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 035_16PEL

A new CV is being submitted with this protocol ☐

An updated CV is being submitted with this protocol ☐

Please note that your ISAC application form and protocol **must be copied to all e-mail addresses*

listed above at the time of submission of your application to the ISAC mailbox. Failure to do so will result in delays in the processing of your application.

11. Conflict of interest statement*

Please provide a draft of the conflict (or competing) of interest (COI) statement that you intend to include in any publication which might result from this work

No relevant conflicts of interests.

**Please refer to the International Committee of Medical Journal Editors (ICMJE) for guidance on what constitutes a COI.*

12. Experience/expertise available

Please complete the following questions to indicate the experience/ expertise available within the team of investigators/collaborators actively involved in the proposed research, including the analysis of data and interpretation of results.

	Previous GPRD/CPRD Studies	**Publications using GPRD/CPRD data**
None	☐	☐
1-3	☐	☐
> 3	☒	☒

Experience/Expertise available	**Yes**	**No**

Is statistical expertise available within the research team? Professor of Medical Statistics Rafael Perera is the Chief Investigator. Sarah Stevens, Toqir Mukhtar and Clare Bankhead are all Primary Care Epidemiology Statisticians with extensive CPRD experience	☒	☐
Is experience of handling large data sets (>1 million records) available within the research team? *If yes, please indicate the name(s) of the relevant investigator(s)* Professor Rafael Perera, Professor Paul Little, Professor Chris Salisbury, Dr Clare Bankhead, Sarah Stevens and Dr Torqi Mukhtar and Professor Richard Hobbs have all undertake and published studies using the CPRD.	☒	☐
Is experience of practising in UK primary care available to or within the research team? *If yes, please indicate the name(s) of the relevant investigator(s)* Professor Carl Heneghan, Professor Paul Little, Professor Chris Salisbury and Professor Richard Hobbs are all UK-based General Practitioners.	☒	☐

13. References relating to your study

Please list up to 3 references (most relevant) relating to your proposed study:

1. Hobbs FDR, Bankhead C, Mukhtar T, Stevens S, Perera-Salazar R, Holt T, et al. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14. Lancet; 2016 Jun;387(10035):2323–30.
2. Mukhtar TK, Molyneux AJ, Hall N, et al. The falling rates of hospital admission, case fatality, and population-based mortality for subarachnoid hemorrhage in England, 1999-2010. J Neurosurg 2016:1-7.
3. NHS Right Care. Diagnostics: The NHS Atlas of Variation in Diagnostic Services. 2012;(November):220.

SECTION C: ACCESS TO THE DATA

14. Financial Sponsor of study[§]

[§]*Please note: The name of the source of funding will be published on CPRD's website as part of its transparency policy*

Pharmaceutical Industry ☐ *Please specify name and country:*

Academia ☒ **NIHR School of Primary Care Research**

Government / NHS ☐ *Please specify name and country:*

Charity ☒ **Primary Care Research Trust, UK**

Other ☐ *Please specify name and country:*

None ☐

15. Type of Institution conducting the research

Pharmaceutical Industry ☐ *Please specify name and country:*

Academia ☒ *Please specify name and country:* University of Oxford, UK

Government Department ☐ *Please specify name and country:*

Research Service Provider ☐ *Please specify name and country:*

NHS ☐ *Please specify name and country:*

Other ☐ *Please specify name and country:*

16. Data access arrangements

The financial sponsor/ collaborator* has a licence for CPRD GOLD and will extract the data

☐

The institution carrying out the analysis has a licence for CPRD GOLD and will extract the data**

☒

A data set will be provided by the CPRD[¥€] ☐

CPRD has been commissioned to extract the data and perform the analyses[€]

☐

Other: ☐

If Other, please specify:

**Collaborators supplying data for this study must be named on the protocol as co-applicants.*

***If data sources other than CPRD GOLD are required, these will be supplied by CPRD*

[¥]*Please note that datasets provided by CPRD are limited in size; applicants should contact CPRD (kc@cprd.com) if a dataset of >300,000 patients is required.*

[€]*Investigators must discuss their request with a member of the CPRD Research team before submitting an ISAC application. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email (kc@cprd.com) to discuss your requirements. Please also state the name of CPRD Research team with whom you have discussed this request (provide the date of discussion and any relevant reference information):*

Name of CPRD Researcher Reference number (where available) Date of contact

17. Primary care data

Please specify which primary care data set(s) are required)

Vision only (Default for CPRD studies) ☒ Both Vision and EMIS®* ☐

EMIS® only* ☐

Note: Vision and EMIS are different practice management systems. CPRD has traditionally collected data from Vision practice. Data collected from EMIS is currently under evaluation prior to wider release.

Investigators requiring the use of EMIS data **must discuss the study with a member of the CPRD Research team before submitting an ISAC application*

Please state the name of the CPRD Researcher with whom you have discussed your request for EMIS data:

Name of CPRD Researcher Reference number (where available) Date of contact

SECTION D: INFORMATION ON DATA LINKAGES

18. Does this protocol seek access to linked data

Yes* ☒ No☐ If No, please move to section E.

We, as a research group, have accessed these CPRD linked data resources and are not seeking HES Accident and Emergency data or HES Diagnostic Imaging Dataset data.

Research groups which have not previously accessed CPRD linked data resources **must discuss access to these resources with a member of the CPRD Research team, before submitting an ISAC application. Investigators requiring access to HES Accident and Emergency data, HES Diagnostic Imaging Dataset and PROMS data **must** also discuss this with a member of the CPRD Research team before submitting an ISAC application. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email kc@cprd.com to discuss your requirements **before** submitting your application.*

Please note that as part of the ISAC review of linkages, your protocol may be shared - in confidence - with a representative of the requested linked data set(s) and summary details may be shared - in confidence - with the Confidentiality Advisory Group of the Health Research Authority.

19. Please select the source(s) of linked data being requested[§]
[§]Please note: This information will be published on the CPRD's website as part of its transparency policy.

☐ ONS Death Registration Data

☐ HES Admitted Patient Care

☐ HES Outpatient

☐ HES Accident and Emergency

☐ HES Diagnostic Imaging Dataset

☐ MINAP (Myocardial Ischaemia National Audit Project)

☐ Cancer Registration Data*

☐ PROMS (Patient Reported Outcomes Measure)**

☐ CPRD Mother Baby Link

☒ Practice Level Index of Multiple Deprivation (Standard)

☐ Practice Level Index of Multiple Deprivation (Bespoke)

☐ Patient Level Index of Multiple Deprivation***

☐ Patient Level Townsend Score ***

☐ Other**** *Please specify:*

**Applicants seeking access to cancer registration data must complete a Cancer Dataset Agreement form (available from CPRD). This should be submitted to the ISAC as an appendix to your protocol. Please also note that applicants seeking access to cancer registry data must provide consent for publication of their study title and study institution on the UK Cancer Registry website.*

***Assessment of the quality of care delivered to NHS patients in England undergoing four procedures: hip replacement, knee replacement, groin hernia and varicose veins. Please note that patient level PROMS data are only accessible by academics*

**** 'Patient level IMD and Townsend scores will not be supplied for the same study*

*****If "Other" is specified, please provide the name of the individual in the CPRD Research team with whom this linkage has been discussed.*

Name of CPRD Researcher Reference number (where available) Date of contact

20. Total number of linked datasets requested <u>including</u> CPRD GOLD

0

Number of linked datasets requested *(practice/ 'patient' level Index of Multiple Deprivation,*

*Townsend Score or the CPRD Mother Baby Link should **not** be included in this count)*

Please note: Where ≥5 linked datasets are requested, approval may be required from the Confidentiality Advisory Group (CAG) to access these data

21. Is linkage to a local[¥] dataset with <1 million patients being requested?

Yes * ☐ No ☒

**If yes, please provide further details:*

[¥] Data from defined geographical areas i.e. non-national datasets.

22. If you have requested one or more linked data sets, please indicate whether the Chief Investigator or any of the collaborators listed in question 5 above, have access to these data in a patient identifiable form (e.g. full date of birth, NHS number, patient post code), or associated with an identifiable patient index.

Yes* ☐ No ☒

** If yes, please provide further details:*

23. Does this study involve linking to patient *identifiable* data (e.g. hold date of birth, NHS number, patient post code) from other sources?

Yes ☐ No ☒

SECTION E: VALIDATION/VERIFICATION

24. Does this protocol describe a purely observational study using CPRD data?

Yes* ☒ No** ☐

** Yes: If you will be using data obtained from the CPRD Group, this study does not require separate ethics approval from an NHS Research Ethics Committee.*

*** No: You may need to seek separate ethics approval from an NHS Research Ethics Committee for this study. The ISAC will provide advice on whether this may be needed.*

25. Does this protocol involve requesting any additional information from GPs?

Yes* ☐ No ☒

** If yes, please indicate what will be required:*

Completion of questionnaires by the GP[ψ] Yes ☐ No ☐

Is the questionnaire a validated instrument? Yes ☐ No ☐

If yes, has permission been obtained to use the instrument? Yes ☐ No ☐

Please provide further information:

Other (please describe)

ψ Any questionnaire for completion by GPs or other health care professional must be approved by ISAC before circulation for completion.

26. Does this study require contact with patients in order for them to complete a questionnaire?

Yes* ☐ No ☒

**Please note that any questionnaire for completion by patients must be approved by ISAC before circulation for completion.*

27. Does this study require contact with patients in order to collect a sample?

Yes* ☐ No ☒

** Please state what will be collected:*

SECTION F: DECLARATION

28. Signature from the Chief Investigator

- I have read the guidance on '***Completion of the ISAC application form***' and '***Contents of CPRD ISAC Research Protocols***' and have understood these;

- I have read the submitted version of this research protocol, including all supporting documents, and confirm that these are accurate.
- I am suitably qualified and experienced to perform and/or supervise the research study proposed.
- I agree to conduct or supervise the study described in accordance with the relevant, current protocol
- I agree to abide by all ethical, legal and scientific guidelines that relate to access and use of CPRD data for research
- I understand that the details provided in sections marked with (§) in the application form and protocol will be published on the CPRD website in line with CPRD's transparency policy.
- I agree to inform the CPRD of the final outcome of the research study: publication, prolonged delay, completion or termination of the study.

Name: Rafael Perera Date: 27/06/17e-Signature (type name): Rafael Perera

PROTOCOL INFORMATION REQUIRED

The following sections below **must** be included in the CPRD ISAC research protocol. Please refer to the guidance on '***Contents of CPRD ISAC Research Protocols***' (www.cprd.com/isac) for more information on how to complete the sections below. Pages should be numbered. All abbreviations must be defined on first use.

Applicants must complete all sections listed below **Sections which do not apply should be completed as '*Not Applicable*'**
A. Study Title[§] [§]*Please note: This information will be published on CPRD's website as part of its transparency policy* Quantifying variation in UK primary care test use: A 15 year retrospective CPRD analysis of temporal and geographical variation in test use.

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

B. Lay Summary (Max. 200 words)[§]

Over the last decade, the NHS has experienced an unprecedented rise in spending; total expenses for the NHS have increased by over £50bn, an 80% rise. Given large numbers of patients are seen in primary care (the majority of NHS consultations take place in primary care), rising expenses threaten the ability of the NHS to provide effective healthcare. A recent study suggests there is an urgent need to better understand GP workload, particularly changes in the number of tests ordered from primary care over time.

To address this, we will determine if the total number of tests ordered from primary care has changed over time (2000 – 2015), the change in use of 25 specific tests over the same time period and differences in test use across different parts of the UK.

This research will determine if test ordering from general practice has increased over time, identify specific tests that have increased most rapidly over time and identify specific tests that are used more frequently in some regions in the UK. These results will inform policy makers of changes in test use, which can help contribute to policy changes required to help NHS sustainability.

C. Technical Summary (Max. 200 words)[§]

[§]*Please note: This information will be published on CPRD's website as part of its transparency policy*

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

Over the last decade, the NHS has experienced an unprecedented spending rise, net expenditure has increased by over £50bn, an ~80% rise. Increasing expenditure threatens NHS sustainability.

Primary care accounts for 90% of NHS contact, but is on the brink of 'saturation point'. A recent Lancet study recommends quantifying primary care test use workload (ISAC Protocol Number: 15_120R). We aim to determine temporal change in total test use (2000 – 2015), temporal change in 25 specific tests and geographical differences in total and specific test use.

Tests ordered for patients from UK general practices registered in the CPRD from April 1st 2000 to March 31st 2015 will be analysed. We will use joinpoint regression to quantify temporal changes in test use. Trends will be examined nationally and also compared across different geographical regions. Using the same methodology we will examine the temporal and geographical variation for 25 specific tests.

This research will facilitate a greater understanding of primary care activity and help shape policy, resource allocation and budgeting. Our results will also lead to further research, particularly investigating explanations for temporal and/or geographical variation. The ultimate aim of this research is to contribute to the changes required to ensure the sustainability of the NHS.

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

D. Objectives, Specific Aims and Rationale

Objective 1: To determine the temporal change in test use in UK primary care from 2000 – 2015

Aim 1.1: To examine changes in total test ordering in UK primary care from 2000-2015

Aim 1.2: To examine changes in use of 25 specific tests in UK primary care from 2000 – 2015 (see Appendix A for lists of specific tests)

Objective 2: To determine geographical differences in test use in UK primary care

Aim 2.1: To determine geographical variation in total test use exists across the UK

Aim 2.2: To quantify geographical variation in use of 25 specific tests exists across the UK

Rationale

Primary care is the cornerstone of effective health care. It remains integral to the NHS, with primary care accounting for ~90% of patient contact (1). Despite its importance, very little research is available on the volume and activity of primary care. Given the current and likely future financial austerity facing healthcare systems around the world, understanding primary care test use workload has never been greater.

General Practitioners provide a broad array of services, including the utilisation of tests, to screen, diagnosis and monitor disease. Although tests are one of the front-line resources GPs

have available to assist them, almost nothing is known about the volume of test use in General Practice.

Examining temporal changes in test use will be essential for UK primary care resource allocation and budgeting. Further, an understanding of trends in total and specific test use over time and between different geographical regions will lead to research exploring reasons for variation and, if relevant, the drivers and mechanisms contributing to increased test use. It will also lead onto to an examination of the appropriateness of these changes in test use. With an understanding of the above, we will be ideally placed to run impactful randomised controlled trials (RCTs) to mitigate geographical variation and improve primary care test use. We expect our results to contribute to the steps required to sustain UK primary care.

Lastly, our results will inform requirements for future GP training (e.g. help determine GP numbers to accommodate projected workload), as well as directly influence clinical practice through alerting GPs to high volume tests and geographical variation in use.

E. Study Background

Over the last decade, the NHS has experienced an unprecedented rise in spending (2,3). Since 2003, net expenditure has increased by over £50bn; an ~80% rise (3). These increasing costs threaten the sustainability of the NHS.

Similar to many healthcare systems, primary care provides the majority of care within the NHS;

accounting for 90% of all contact (1). Hobbs et al (1) recently reported that UK General Practice clinical workload has markedly increased and is on the brink of 'saturation point' (1) (ISAC protocol number: 15_120R). They reported a marked lack of research exploring primary care activity and workload and recommend, to help alleviate clinical workload burden, further research to quantify and understand primary care activity and volume, particularly 'investigation trends' (1).

Reaching a diagnosis in primary care is exceedingly complex (4). The combination of undifferentiated symptoms, a low prevalence of serious disease, a high degree of symptom overlap between serious and benign conditions and patients with multiple complaints all complicate reaching a diagnosis (4). In around 40% of GP consultations, a diagnosis is unable to be established from history and physical examination alone (5). Tests are often used to help clinicians reach a diagnosis (4,6).

Greater access to tests (7), medicolegal consequences of under testing (8), clinical performance measures (9) and increasing time constraints (9) are all drivers of increasing test use. In the US, 81% of clinicians admit to increasing their test use due to medico-legal fears (10) and almost half of all primary care physicians believe 'that their own patients are receiving too much care' (9).

Despite the volume of care provided from primary care (90% of all NHS contact) and the approaching workload 'saturation point', there is a paucity of data regarding primary care activity

Applicants must complete all sections listed below Sections which do not apply should be completed as '*Not Applicable*'
and resource use. Further, despite documented drivers and physician concerns about increasing and unnecessary testing in primary care there is no comprehensive data quantifying test trends or drivers of increased test use. To meet this need, this study aims to quantify temporal changes in test use from primary care over the last 15 years and quantify geographical variation in test use.
F. Study Type Our study will descriptively explore changes in test ordering over a) time and b) space (geography). Objective 1 (*Aim 1.1 and 1.2):* Descriptive Explore temporal changes in test ordering from UK General Practice. Objective 2 *(Aim 2.1 and 2.2):* Descriptive A descriptive comparison in total and specific test use across different regions in the UK
G. Study Design To achieve our objectives we will conduct an observational study employing a retrospective, longitudinal design (Objective 1: Temporal trends) and a cross-sectional analysis (Objective 2: Geographical variation). For Objective 1, we will determine temporal trends using joinpoint regression and for Objective 2, we will descriptively compare geographical variation in total and specific test use using a cross sectional design.
H. Feasibility counts As our study will investigate the temporal change in total test use per person years from 2000-2015. As such, our unit of analysis will be tests. Members of the CPRD scientific committee for the University of Oxford have stated that, as of July 2016, the total number of test results available within CPRD is 1,032,989,274 from 10,232,800 patients. Our study aims to produce a

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

test ordering rate per person years for each year (2000 – 2015) and for each geographical region (across 12 CPRD geographical regions) for 2015.

I. Sample size considerations

As our study aims to quantify and describe changes and variation in total and specific test ordering, we will require a sample that captures all tests over the last 15 years. The CPRD scientific committee for the Department of Primary Care Health Sciences, University of Oxford, has advised that as of July 2016, the total number of historical test results records is 1,032,989,274 from 10,232,800 patients. We anticipate a sample of this size to provide total and specific test ordering rate per year, and per geographical region with very tight confidence intervals. To illustrate this, we have estimated the rate of test ordering per person years for one of the analysis we anticipate to have a relatively small sample size. We anticipate the rate of ordering for one of the 25 specific tests in a specific geographical region to have the smallest sample sizes. We anticipate, however, that these analyses will still have very tight confidence intervals. Below we have calculated a crude estimate of the rate of chest x-ray ordering in 2015 with corresponding confidence intervals (1), as well as the rate of chest x-ray ordering in 2015 for one of the 12 geographical regions (2).

1. <u>Chest x-rays per person year for 2015</u>

a) Total number of CXR test events (Read Database): 2,645,756
b) Period of data: 29 years (1987 – 2016) (11)
c) Average number of chest x-rays per year (a/b): 91,233
d) Average number of patients in CPRD per year: 352,855 (10,232,800/29)
e) Average number of CXRs per person year (c/d): 0.2586 (95%CI: 0.2571 to 0.2600) chest x-rays per person year

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

2. Chest x-rays per person year 2015 for one geographical region

f) Average number of CXRs per year per region: (c/12): 7603 (91,233/12)
g) Average number of patients per region (d/12): 29405 (352855/12)
h) Average number of CXRs per person year per region: 0.2586 (95%CI: 0.2536 to 0.2636)

J. Data Linkage Required (if applicable):[§]

[§]*Please note that the data linkage/s requested in research protocols will be published by the CPRD as part of its transparency policy*

The Practice Level Index of Multiple Deprivation linkage is being requested. This data linkage will be used to standardise our analysis of test ordering rates across different geographical regions. Standarisation via deprivation data will adjust for any confounding caused by geographical differences in deprivation. 2015 deprivation data is only available in approximately 75% of English General Practices. We will perform two sensitivity analyses on this 75% subset English practices (one without and one with adjustment for deprivation) and compare these results with our analysis across the whole of the UK. See section N for further details.

K. Study population

We will analyse total tests ordered for patients from UK general practices registered in the CPRD from April 1st 2000 to March 31st 2015. We will include all tests ordered from a General Practitioner in Primary Care regardless of the setting the test was performed in. For example, an endoscopy ordered by a GP, but performed in secondary care will be included. However, an endoscopy ordered by a Gastroenterologist and performed in the same secondary care setting will be excluded. We will include radiological, laboratory, endoscopic (i.e. colonoscopy) and physiological (exercise stress test and spirometry) tests. Total tests is defined as the total

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

number of all tests ordered by a General Practitioner, excluding physical examination findings, vital signs and measurement of weight. Tests unlikely to be ordered by GPs will also be excluded to reduce the risk of tests ordered in secondary care being included in our study in error. These excluded tests have been derived from previous similar studies using CPRD (12), they include: Uric acid blood level, serum chloride, serum bicarbonate, anion gap, blood gases, ascitic fluid examination, Schilling test. Thus, 'all tests/total tests' will include the tests listed in Appendix A and B as well as all other tests registered as ordered in CPRD over the selected dates (for instance, an x-ray of the wrist would be included amongst many others).Test data will only be included for the period of time if the practice is up to standard and the patient record is acceptable for research. Our analysis of test use will be calculated and presented as a ratio per person years. We will calculate our denominator (person years) by calculating the total number of person years patients are registered across all General Practices for each year. For example, when patients are registered for less than a year, they will be included, but their contribution to the person year calculation will be adjusted, such that a patient that was only registered for 6 months would contribute 0.5 person years. This figure will then be age and sex standardised to the mid-2015 English population (15).

L. Selection of comparison group(s) or controls

Aim 1.1: No comparison group is required

Aim 1.2 and Aim 2.2: Temporal and geographical change of 25 specific tests will be analysed (see Appendix A). These specific tests were selected if they met one of two criteria: 1. Specific guidance on their use in primary care is stated in one or more of the following guidelines: QOF, NICE, Choosing Wisely and/or NICE Do No Do Guidelines or 2. Recognised as common tests ordered from primary care, where common is one of the two most frequent laboratory tests or radiological tests ordered from Oxfordshire Primary Care 2010 – 2015 (data analysed in a separate, on-going study). Geographical regions selected for comparison are London, North East, North West, Yorkshire, West Midlands, East Midlands, East, South West, South East, Wales, Scotland, and Northern Ireland, which are the predefined regions within CPRD.

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

Aim 2.1: Geographical regions for descriptive comparison are London, North East, North West, Yorkshire, West Midlands, East Midlands, East, South West, South East, Wales, Scotland, and Northern Ireland, as stated above.

M. Exposures, Health Outcomes[§] and Covariates

[§]*Please note: Summary information on health outcomes (as included on the ISAC application form above)will be published on CPRD's website as part of its transparency policy*

Outcomes

Aim 1.1:

Outcome: Rate of total test ordering per person-years for each year (2000 – 2015) from all UK General Practices.

Aim 1.2:

Outcome: Rate of test ordering for each specific test (Appendix A) per persons-years for each specific year (2000 – 2015) from all UK General Practices.

Aim 2.1:

Outcome: Rate of total test ordering per person-years for each geographical region in 2015.

Aim 2.2:

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

Outcome: Rate of each specific test ordering per person-years for each geographical region in 2015.

Operational definitions

Total tests: All tests ordered by General Practitioners except physical examination findings, vital signs, measurement of weight, uric acid blood level, serum chloride, serum bicarbonate, anion gap, blood gases, ascetic fluid examination and Schilling test. This definition is further defined in Section K Study Population section.

The CPRD releases a monthly update of the total count of tests ordered, this data for each year from April 1st 2000 to March 31st 2015 will be sought and stratified per year (e.g. April 1st 2000 – March 31st 2001).

'Geographical regions' are defined by as they are in the CPRD: London, North East, North West, Yorkshire, West Midlands, East Midlands, East, South West, South East, Wales, Scotland, and Northern Ireland.

N. Data/ Statistical Analysis

The foundation of our statistical analysis will utilise joinpoint regression (13). Joinpoint regression models consists of straight lines which are connected by joinpoints (also referred to as knots), and each joinpoint is the estimated location of a change in the slope of the trend line. The joinpoint programme starts from a null hypothesis model of zero joinpoints and tests whether the alternative hypothesis of the maximum number of joinpoints specified for the model has a

statistically significant lower sum of residual squares (14). Permutation tests are applied sequentially until a model of best fit is reached (15). Because multiple tests are performed, the significance level of each test is adjusted to control the overall type I error at specified a level (0.05) (16). The maximum number of joinpoints is dependent upon the number of datapoints in the model, and for 15 datapoints, this would be 2 joinpoints.

Aim 1.1: To examine changes in total test ordering in UK primary care from 2000-2015

To address our first objective the total number of tests ordered from registered practices will be determined for each year (2000-2015), e.g. from April 1st 2000 to March 31st 2001. Concurrently, the total number of patients registered across all General Practices for each year will be determined and age and sex standardised to the mid-2015 English population (17). The total number of person years for each year will then be calculated and a rate of total test ordering per person-years for each year will be generated. When patients are registered for less than a year, they will be included, but their contribution to the person year calculation will be adjusted. For instance, a patient that was only registered for 6 months would contribute 0.5 person years. For comparison across years, joinpoint regression will then be used to model trends over time. The annual percentage change between joinpoints will then be calculated using significance level of α=0·05 (Monte Carlo permutation test) with corresponding confidence intervals. Where possible, results will also be presented stratified by type of test, e.g. radiological, laboratory or other (endoscopy, spirometry etc) and key guidelines/QOF changes will be labelled.

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

Aim 1.2: To examine changes in use of 25 specific tests in UK primary care from 2000 – 2015

Using a similar methodology to Aim 1.2, the total number of each specific test (Appendix A) ordered will be determined for each year (2000 – 2015), followed by calculating a rate of test ordering per-person years. The change in total number of each specific tests ordered will then be modelled over time using joinpoint regression. Each of the 25 tests (Appendix A) will be analysed separately and the results of each test will be descriptively compared against the national average. These results will be displayed graphically with confidence intervals surrounding the national average and individual test rates overlaid, key guidelines/QOF changes will be labelled on the graphs. The five tests with the greatest discrepancy from the national average will be identified. Future studies will focus on exploring this variation to determine its appropriateness. The annual percentage change between joinpoints will then be calculated using significance level of $\alpha=0{\cdot}05$ (Monte Carlo permutation test) with corresponding confidence intervals. Where significant change occurs over time, we will explore explanations for this, for instance change in guidance, new evidence suggesting an improvement or worsening in a tests performance, performance related incentives to order etc.

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

Aim 2.1: To determine if geographical variation in total test use exists across the UK

The total number of all tests ordered per geographical region (London, North East, North West, Yorkshire, West Midlands, East Midlands, East, South West, South East, Wales, Scotland, and Northern Ireland) for 2015 will be calculated. The total person-years of registered patients from each geographical region for 2015 will then be calculated and age and sex standardised to the mid 2015 English population (17). The rate of total test use from each geographical area will then be calculated (total number of tests per person-years) with corresponding 95% confidence intervals. This total test rate for each geographical region will be compared against the national test rate for 2015 (calculated from Aim 1.1). Results from geographical regions that are above or below the upper and lower 95% confidence intervals of the national test rate will be considered significantly different from the national test rate. Further, results between geographical regions without overlapping confidence intervals will be considered significantly different from each other.

After the above primary analysis is conducted, we will conduct sensitivity analyses to control for the potential confounding of geographical differences in patient deprivation. We will perform two sensitivity analyses. The first will explore geographical variation in total test use including only the 75% of English General Practices that have deprivation linkage, but this analysis will not be adjusted for deprivation. The results from this analysis will be compared to the primary analysis (geographical variation of total test use in all UK General Practices). The third analysis will include the same 75% of English General Practices, but will be controlled for deprivation. The

Applicants must complete all sections listed below Sections which do not apply should be completed as '*Not Applicable*'
results from these three analyses will be compared and differences explored. *Aim 2.2:* To quantify geographical variation in specific test use across the UK The total number of each specific test (Appendix A) ordered per geographical region for 2015 will be calculated. The total number of registered patients from each geographical region for 2015 will then be calculated and age and sex standardised to the mid 2015 English population (12). The rate of test use for each specific test for each geographical area will then be calculated (total number of tests per person-years) with corresponding 95% confidence intervals. Variation between geographical regions for each specific test will be quantified via comparison to the national average. Order rates for specific tests will also be compared between each geographical region. For instance the test rate of chest x-ray ordering for 2015 in East Midlands will be compared against chest x-ray ordering from the 11 other geographical regions, as well as the national chest x-ray ordering rate (calculated in Aim 1.2). Point estimates and associated confidence intervals will also be presented graphically for visual comparison with national estimates. As described in Aim 2.1, two sensitivity analyses will be performed to control for differences in deprivation. Where substantial geographical variation exists, we will explore explanations for this, for instance change in guidance, new evidence suggesting an improvement or worsening in a tests performance, performance related incentives to order etc.
O. Plan for addressing confounding The three factors that could confound our results are age, gender and differences in population

characteristics, either over time or over different geographical regions.

We plan to address these potential confounding factors by:

1. Age and sex standardising the total number of patients identified for each year (in regards to Aim 1) and for each geographical region (in regards to Aim 2). Standardisation will occur to the mid 2015 English population (17). This standardisation will account for differences in patient age and/or gender between different years (Aim 1) and different geographical regions (Aim 2).
2. Test ordering rate per person years: Differences in total patient numbers across a) years (Aim 1) and b) geographical regions (Aim 2) will be accounted for by reporting the number of tests ordered as a rate of person years. This standardisation method is described in Section N and M.

For Aim 2 of our analysis, we will also standardise for deprivation (where available). Given the association between deprivation and co-morbidities (19), standardising our analysis to adjust for deprivation will not only account for deprivation differences across years (Aim 1) and geographical regions (Aim 2), but also account for differences in co-morbidities.

P. Plans for addressing missing data

Age and sex co-variants, which we will use to standardise our analysis, are fully observed in the CPRD dataset. Deprivation, the remaining co-variant we will use to standardise our analysis, is fully observed for approximately 75% of English General Practices. To overcome missing data in Northern Ireland, Scotland and Wales, we will performed sensitivity analyses on this 75% subset of English General Practices and compare results to our main results (as described in Section N Data/Statistical Analysis).

It is plausible that we will miss tests that have been ordered by General Practitioners but only

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

recorded in free text. However, given our study over starts after 2000, the quality of coded CPRD data is more robust and electronic health were implemented across British General Practices. This will reduce the risk of missing free text entries because test ordering data is automatically entered into EHRs and thus their test ordering will be captured in CPRD code. Therefore, we expect this missing data to be minimal. If substantial data is written as free text and missing, our results will under-estimate changes in test ordering.

Q. Patient or user group involvement (if applicable)

This PPIE summary was produced with Lynne Maddocks (LM), the Department Primary Care's PPIE coordinator.

Our PPIE strategy is summarised below, stratified by type and timing of PPIE

Timing of PPIE

1. **Pre-data analysis:** Four 1-hour interviews with members of the public have been conducted. These members of the public were from varied backgrounds with varied co-morbidities (cardiovascular, mental health, endocrinological, gynaecological and gastrointestinal). The interviews were semi-structured and covered edits of the Plain English Summary, patient's thoughts on importance and relevance of research and a series of themed questions around patient's experience with GP test ordering. From these interviews, the following themes emerged: The following themes emerged:
    - Greater use and reliance on tests: All patients stated that they feel test use is increasing.
    - Patient desire for test use: 3 out of 4 participants reported multiple situations where they thought GPs should use tests more, because, as one patient put it 'the more information about our bodies, the better'. They reported they often had to 'beg' for a test.
    - Knock on consequences of testing: All patients reported follow up requirements after a test. Some of these were positive (e.g. referral to a specialist for confirmation of a diagnosis) and some were negative (a series of screening tests and outpatient

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

appointments, for instance one participant was referred to a 'diabetic depression screening clinic' after an elevated HbA1c (and diagnosis of diabetes). All participants noted the knock on resource implications from testing.

2. **Once data analysis complete:** Upon completion of data analysis, we will conduct a half-day patient advisory group workshop with 4-6 participants. These patients will be reimbursed for their time, including a contribution to travel costs. We will discuss results from our analysis and set out to achieve the following
    a. Explore variation and generate hypotheses as to why certain variation exists, which will guide follow up studies (and potentially other studies) across our four aims (temporal changes in total test use, geographical variation in total test use, temporal variation of any of the 25 selected tests, geographical variation of any of the 25 selected tests)
    b. Determine how results can be used to better inform GP resource requirements, patient's thoughts on other NHS workload changes and the value in conducting further research to explore this burden, for instance, prescribing changes
    c. Generate dissemination and presentation of findings strategies, including Plain English Summaries.
3. **After completion of study:** Once the study is complete, we will distribute lay summaries via: Twitter (~10,000 followers accumulatively), the Department website, Patient Participation Groups (PPGs) within Oxfordshire General Practices and via local Clinical Commissioning Groups (CCGs) (Oxford, Bristol and Southampton). We will also take any advice/guidance generated from the patient advisory group.

Type of PPIE

In accordance with the INVOLVE national advisory group Framework for 'Public Involvement in

research' our activities will be defined under the following headings:

Involvement

- Four members of the public have been interviewed as stated above
- The Patient Advisory Group workshop (above) will give members of the public an active role in this study, interpreting results, designing follow up studies, producing lay summaries and dissemination strategies.

Participation

- The CPRD is a database of prospectively collected patient data. This collection has taken place and my study will revolve around the analysis of this data.

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

Engagement

- Results of this study will be discussed with members of the public and help produce lay summaries. Lay summaries will be distributed widely.

R. Plans for disseminating and communicating study results, including the presence or absence of any restrictions on the extent and timing of publication

Academic publications: 2 academic publications are expected from this work (1. Temporal change (Objective 1) and 2. Geographical variation (Objective 2). In our grant application, funds have been allocated to ensure these studies are available open access. We plan to submit these to publication in March 2017.

Conference presentations: Results from two academic publications will be submitted to national and international conferences (e.g. Evidence Live, Preventing Overdiagnosis).

Wider dissemination: We plan to liaise with the Department of Primary Care Health Sciences and the University of Oxford's media teams to disseminate our results. We hope the established media links of the Department and University will result in print and radio exposure. We also plan to disseminate our findings by writing in publications such as The Conversation and blogs (BMJ, PloS, CEBM, Department). All exposure will be distributed via social media channels (Twitter: >10,000 followers accumulatively).

S. Limitations of the study design, data sources, and analytic methods

Our descriptive study design will allow quantification of temporal and geographical changes in UK General Practice test ordering. Our results will not draw conclusions about why changes have occurred or if they are appropriate. Lack of deprivation data outside England and thus the

inability to standardise for deprivation in Scotland, Wales and Northern Ireland could be considered a limitation of Objective 2 (Aims 2.1 and 2.2). We will account for this by performing a sensitivity analysis with exclusively English General Practices and compare these results to our main results.

The potential of missing tests ordered but only recorded in free text is a potential limitation of our data source, however we have attempted to overcome this but selecting dates where CPRD data quality is more robust (post 2000). We will also counter this issue by considering the 'ordering of a test' as either a) an electronic record indicating the test has been ordered or b) evidence of the result of the test (via result letter for instance). There is evidence to suggest that letter correspondence is incorporated frequently into the CPRD, with around 90% compliance (20,21). Further, the quality and completeness of data in CPRD has been validated extensively and systematically (11,22,23) including validation of the completion of diagnostic records (20,24).

It is plausible that tests ordered may be counted more than once. This may occur because a code may be entered when a test is ordered and another code may be entered when the test result returns to a General Practitioner. We will ensure that tests are not counted more than once by comparing the event date and the system date for each test entry into CPRD. If these dates are different, we will consider the test as entered into CPRD code twice and we will only count the test ordered by the event date. Regardless, if duplication is common these errors should be

consistent over time and space and not affect our core results: assessing change over time and geographical variation.

Lastly, it must be noted that the CPRD reflects around only 7% of UK primary care (11). Thus, although this data is broadly representative of the UK (11,25), it remains a subset. Variation that we anticipate to find from our analyses may be more or less pronounced if data included the entirety of UK primary care. However, given the large size of our sample and the representativeness of the data, we anticipate our results to be valid and reflective of UK primary care test use.

T. References

1. Hobbs FDR, Bankhead C, Mukhtar T, Stevens S, Perera-Salazar R, Holt T, et al. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14. Lancet [Internet]. Elsevier; 2016 Jun [cited 2016 Aug 19];387(10035):2323–30. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0140673616006206

2. Lafond S, Arora S, Charlesworth A, Mckeon A. An analysis of NHS expenditure between 2010 and 2014. 2014;

3. NHS Confederation. Key statistics on the NHS - NHS Confederation [Internet]. Resources. 2016 [cited 2016 Aug 19]. Available from: http://www.nhsconfed.org/resources/key-statistics-on-the-nhs

4. Foot C, Naylor C, Imison C. The quality of GP diagnosis and referral [Internet]. London,

UK: The King's Fund. 2010. Available from: http://amapro.isabelhealthcare.com/pdf/Kings_Fund_Diagnosis_and_Referral_2010.pdf

5. Koch H, van Bokhoven MA, ter Riet G, van Alphen-Jager JT, van der Weijden T, Dinant G-J, et al. Ordering blood tests for patients with unexplained fatigue in general practice: what does it yield? Results of the VAMPIRE trial. Br J Gen Pract [Internet]. Royal College of General Practitioners; 2009 Apr [cited 2016 Nov 9];59(561):e93-100. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19341544

6. Heneghan C, Glasziou P, Thompson M, Rose P, Balla J, Lasserson D, et al. Diagnostic strategies used in primary care. BMJ. 2009;338.

7. Health D of. NHS 2010–2015: from good to great. preventative, people-centred, productive. London; 2009.

8. Esmail A, Neale G, Elstein M, Firth-Cozens J, Davy C VC. Case Studies in Litigation: Claims reviews in four specialties. Manchester; 2004.

9. Sirovich BE, Woloshin S, Schwartz LM. Too Little? Too Much? Primary care physicians' views on US health care: a brief report. Arch Intern Med [Internet]. 2011 Sep 26 [cited 2016 Nov 11];171(17):1582–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21949169

10. Kanzaria HK, Hoffman JR, Probst MA, Caloyeras JP, Berry SH, Brook RH. Emergency Physician Perceptions of Medically Unnecessary Advanced Diagnostic Imaging. Carpenter CR, editor. Acad Emerg Med [Internet]. 2015 Apr [cited 2016 Nov

11];22(4):390–8. Available from: http://doi.wiley.com/10.1111/acem.12625

11. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol [Internet]. Oxford University Press; 2015 Jun [cited 2016 Aug 21];44(3):827–36. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26050254

12. Busby J, Schroeder K, Woltersdorf W, Sterne JAC, Ben-Shlomo Y, Hay A, et al. Temporal growth and geographic variation in the use of laboratory tests by NHS general practices: Using routine data to identify research priorities. Br J Gen Pract. 2013;63(609):256–66.

13. Institute NC. Joinpoint Regression Program. Statistical Methodology and Applications Branch, Surveillance Research Program, National Cancer Institute.;

14. Mukhtar TK, Yeates DR, Goldacre MJ. Breast cancer mortality trends in England and the assessment of the effectiveness of mammography screening: population-based study. J R Soc Med [Internet]. 2013;106(6):234–42. Available from: http://journals.sagepub.com/doi/10.1177/0141076813486779

15. Kim H-J, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joipoint regression with applications to cancer rates. Stat Med. 2000;19(April 1998):335–51.

16. Institute NC. Joinpoint Help Manual, Version 4.5.0.1.

17. Office for National Statistics. Population Estimates for UK, England and Wales, Scotland and Northern Ireland: mid- 2015. 2015;

18. NHS Right Care. The 2nd Atlas of Variation in NHS Diagnostic Services in England.

2017;(January).

19. Charlton J, Rudisill C, Bhattarai N, Gulliford M. Impact of deprivation on occurrence, outcomes and health care costs of people with multiple morbidity. J Health Serv Res Policy [Internet]. 2013;18(4):215–23. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3808175&tool=pmcentrez&rendertype=abstract

20. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: A systematic review. Br J Gen Pract. 2010;60(572):199–206.

21. Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. BMJ [Internet]. 1991;302(6779):766–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1669537&tool=pmcentrez&rendertype=abstract

22. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. BMJ [Internet]. 2003;326(7398):1070. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12750210%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC155692

23. Williams T, van Staa T, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. Ther Adv drug Saf [Internet]. SAGE Publications; 2012 Apr [cited 2016 Aug 21];3(2):89–99. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25083228

Applicants must complete all sections listed below

Sections which do not apply should be completed as '*Not Applicable*'

24. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: A systematic review. Br J Clin Pharmacol. 2010;69(1):4–14.

25. Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, vanStaa T, Grundy E, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. J Public Health (Oxf). 2014;36(4):684–92.

List of Appendices *(Submit all appendices as separate documents to this application)*

Appendix A

Laboratory

1. Cholesterol
2. Thyroid function tests (TFT)
3. Urine albumin/protein:creatinine ratio
4. HbA1c
5. Creatine Kinase (CK)
6. Vitamin D
7. PSA
8. CRP
9. ESR
10. Creatinine

Radiology

1. Pelvic U/S
2. DXA scan
3. Echocardiogram TTE
4. Brain CT
5. Brain MRI
6. Pelvic CT
7. L-spine x-ray
8. Chest x-ray
9. X-ray knee

10. MRI knee
11. MRI L-spine

Invasive

1. Upper endoscopy
2. Lower endoscopy (colonoscopy)
3. Pap smear

Physiological

1. Pulmonary function tests/spirometry

Appendix B

Entity codes

Cholesterol

- 163 Serum Cholesterol

Thyroid function tests

- 203 Thyroid stimulating hormone
- 236 Thyroid function

Urine albumin:creatinine ratio

- 340 Urine biochemistry

HbA1c

- 275 HbA1c Diabetic control

Creatinine Kinase (CK)

Applicants must complete all sections listed below Sections which do not apply should be completed as '*Not Applicable*'
• 164 Creatine phosphokinase *Vitamin D* • 336 Blood trace elements/vitamins *PSA* • 276 Prostate specific antigen *CRP* • 280 C Reactive protein ESR • 273 Erythrocyte sedimentation rate Serum Creatinine • 165 Serum creatinine • 166 Creatinine clearance Pelvic U/S • 238 Pelvic Ultrasound DXA scan

Applicants must complete all sections listed below Sections which do not apply should be completed as '*Not Applicable*'
• 474 DEXA scan/bone mineral density Echocardiogram TTE • 342 Echocardiogram Brain CT • 299 CAT scan • 444 tomography Brain MRI • 300 MRI scan Pelvic CT • 299 CAT Scan Lspine x-ray • 267 spine x-ray CXR • 253 Chest x-ray X-ray knee

- 260 knee x-ray

MRI knee

- 300 MRI scan

MRI L-spine

- 300 MRI scan

Upper endoscopy

- 218 : endoscopy

Lower endoscopy (colonoscopy)

- 218 : endoscopy

Pap smear

- 242 Cervical swab

Pulmonary function tests/spirometry

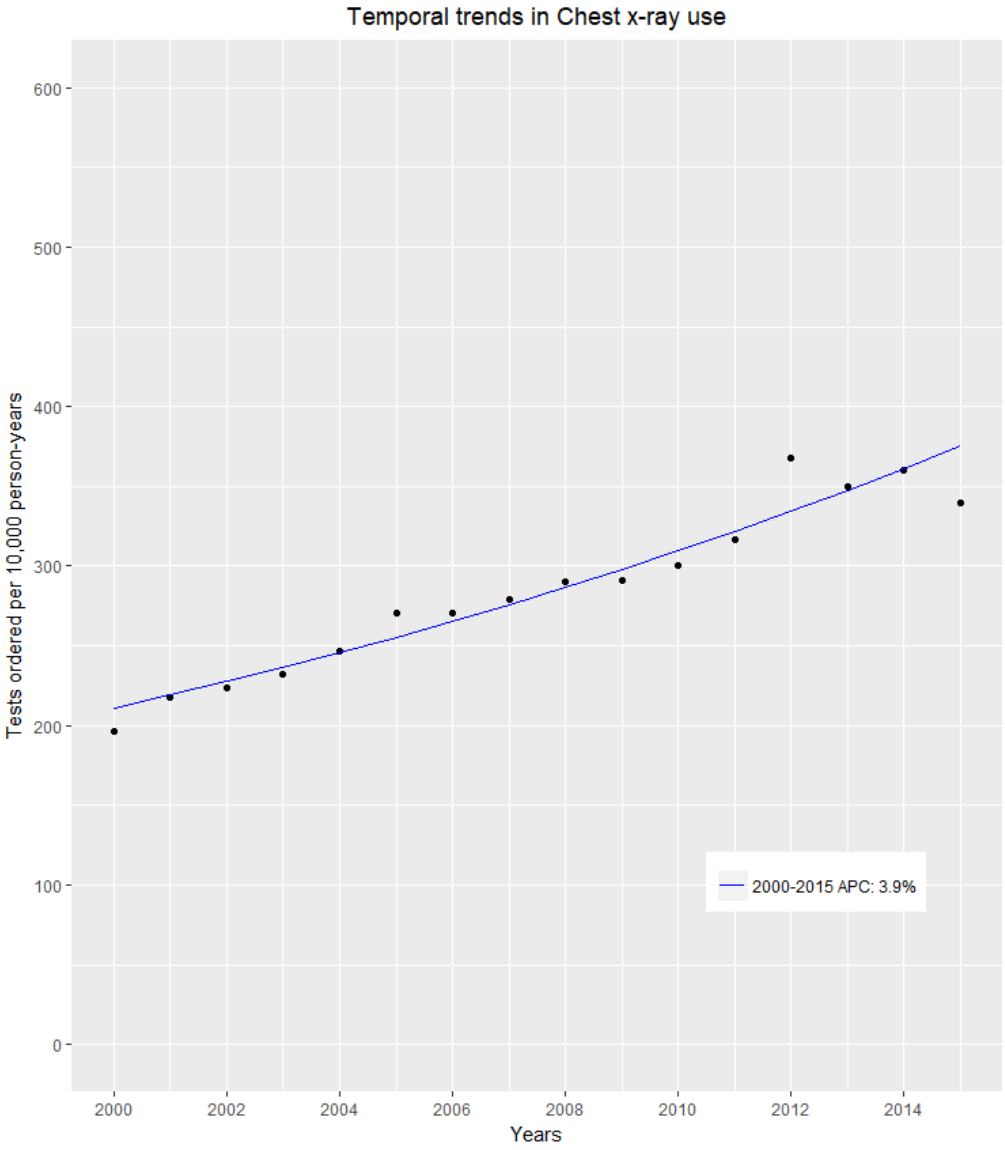
- 380 Pulmonary function tests
- 471 spirometry

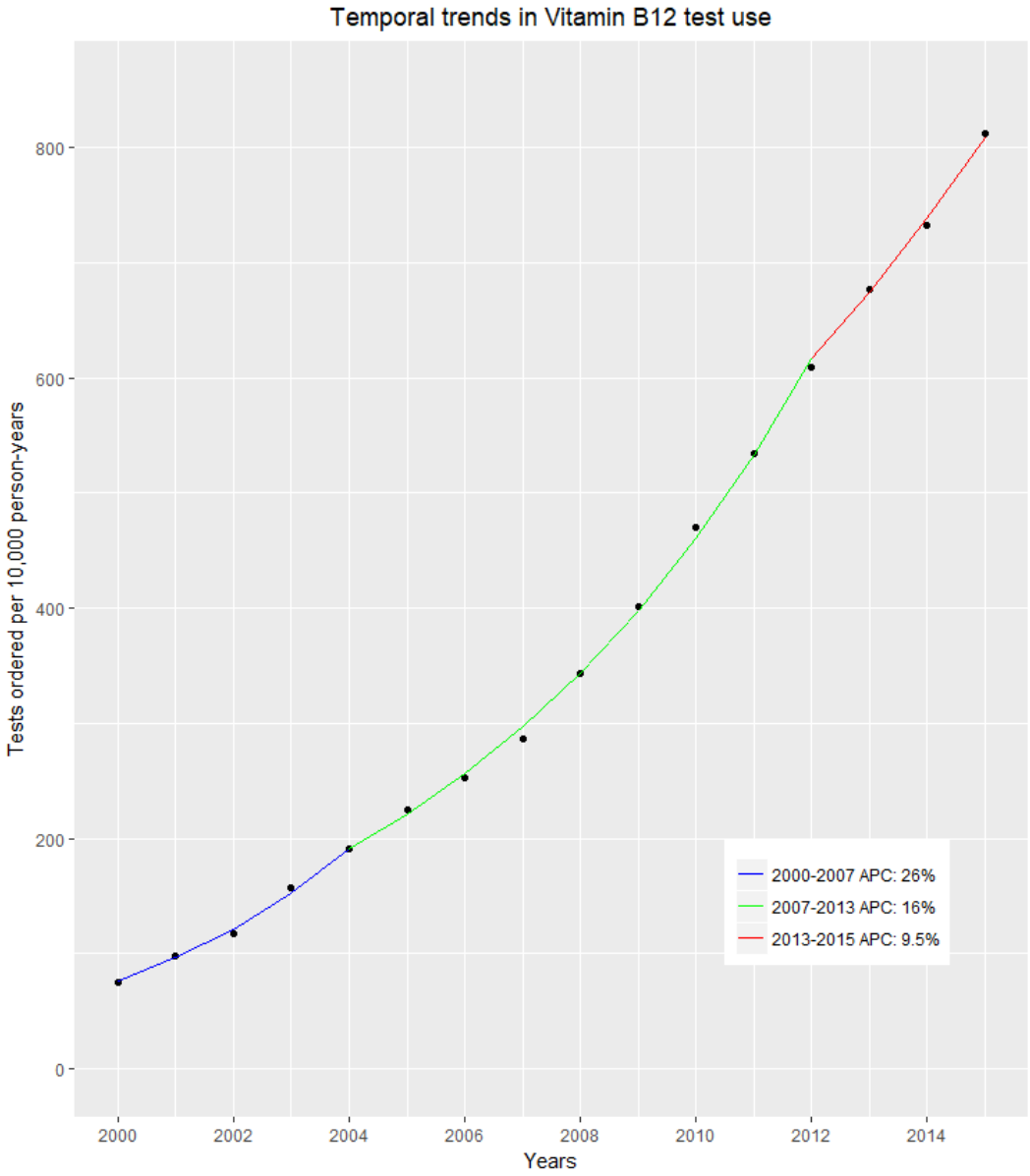
Appendix C

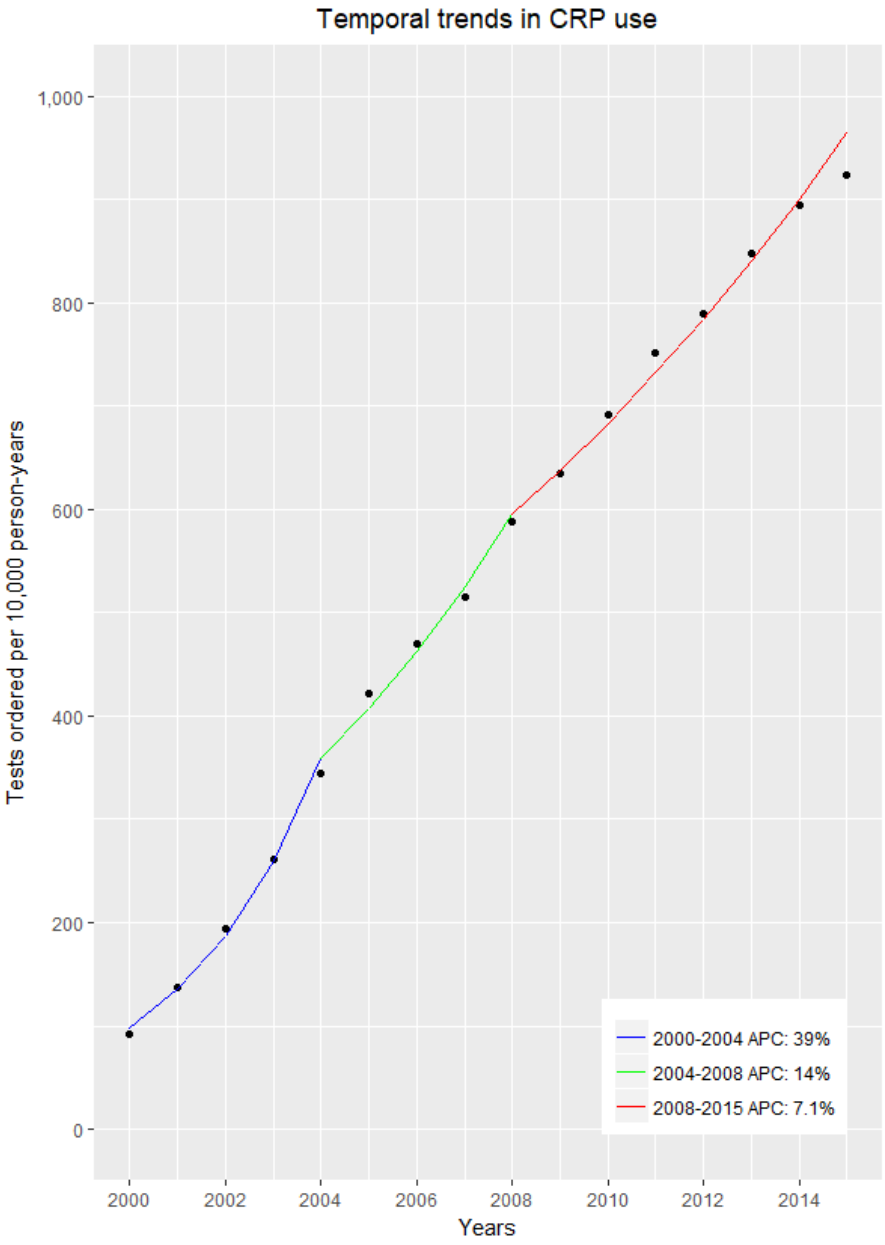
Applicants must complete all sections listed below Sections which do not apply should be completed as '*Not Applicable*'
Attached in separate document

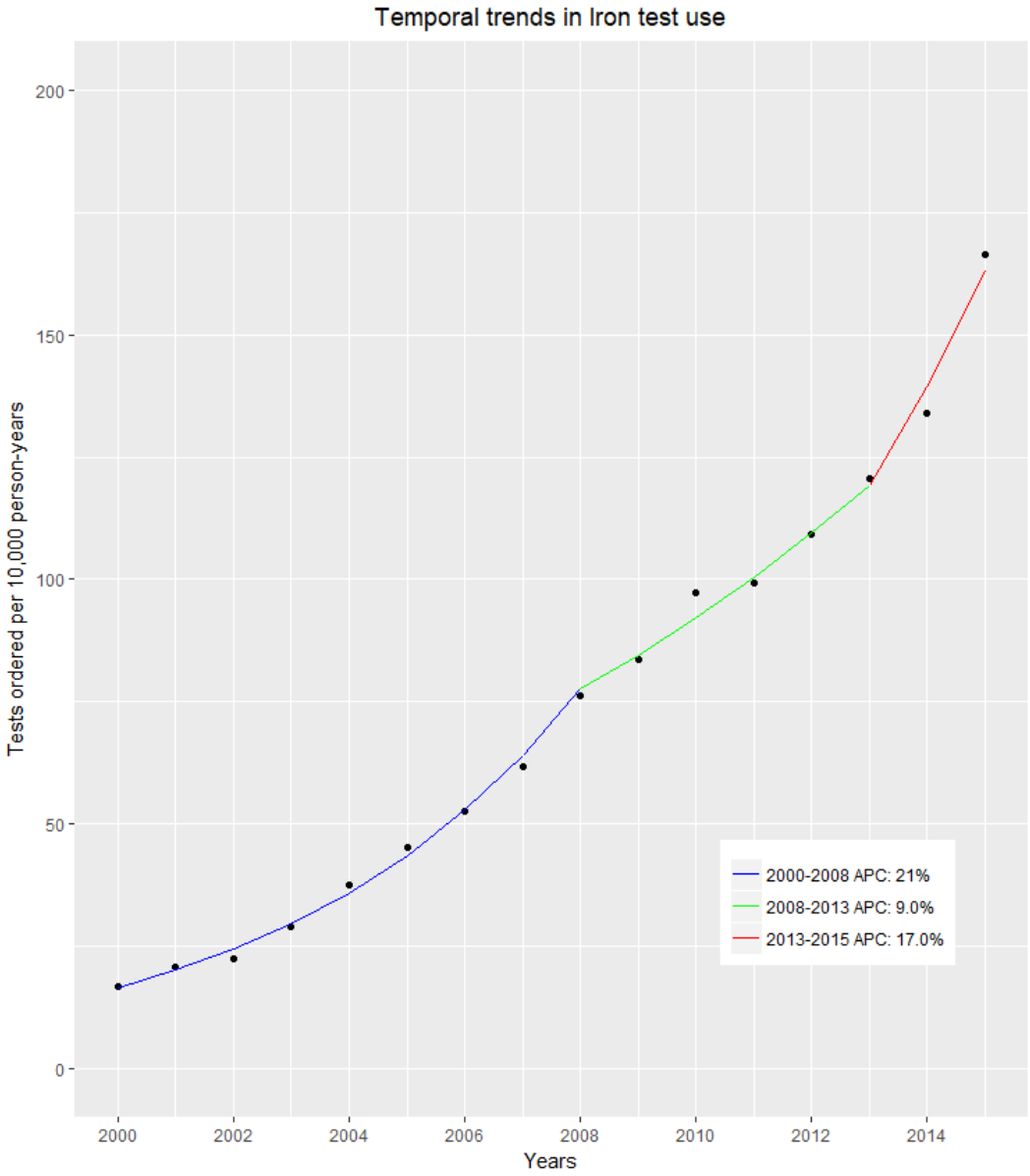
3.2 Temporal patterns of use

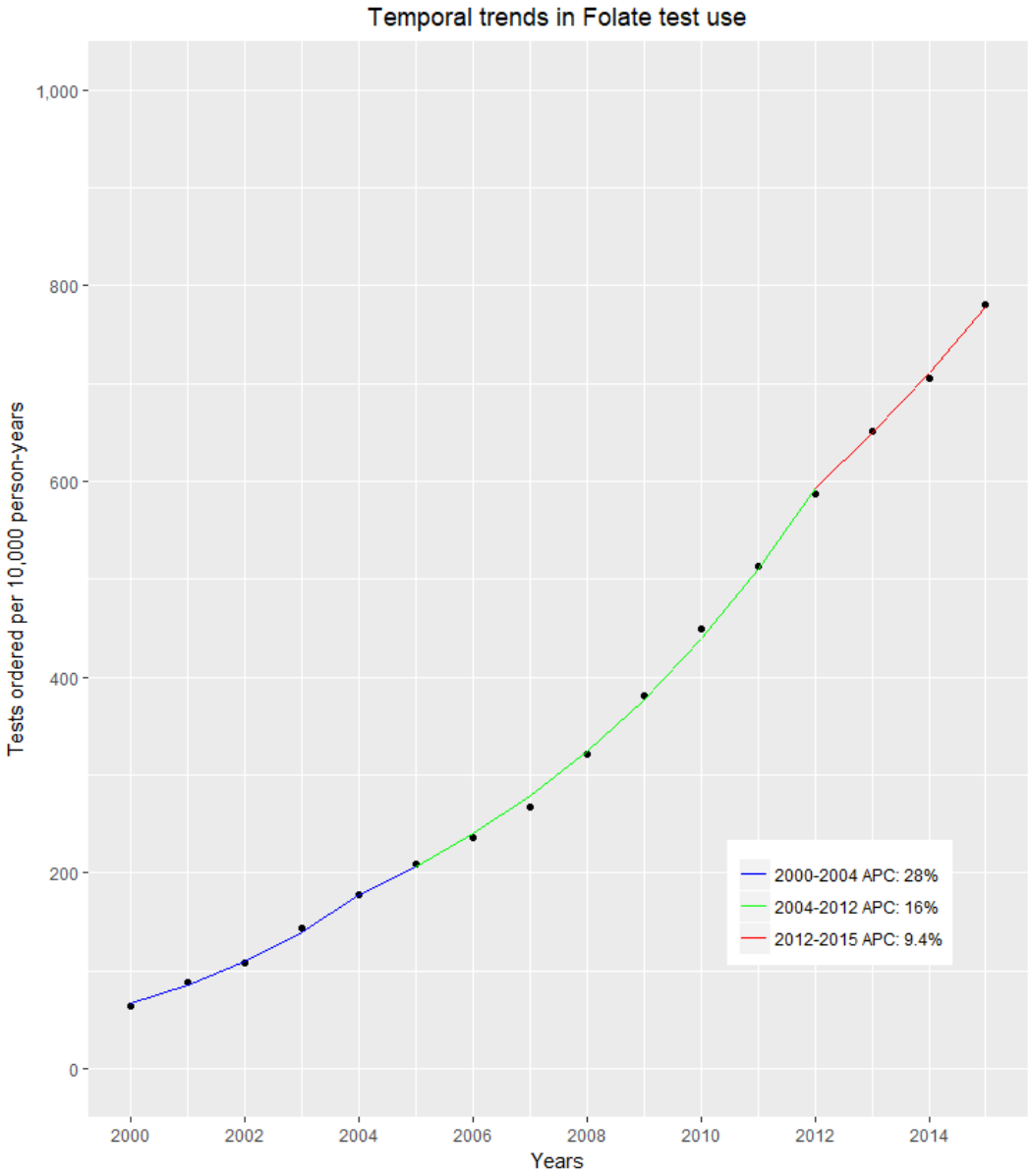
Pattern: Consistent, linear increase

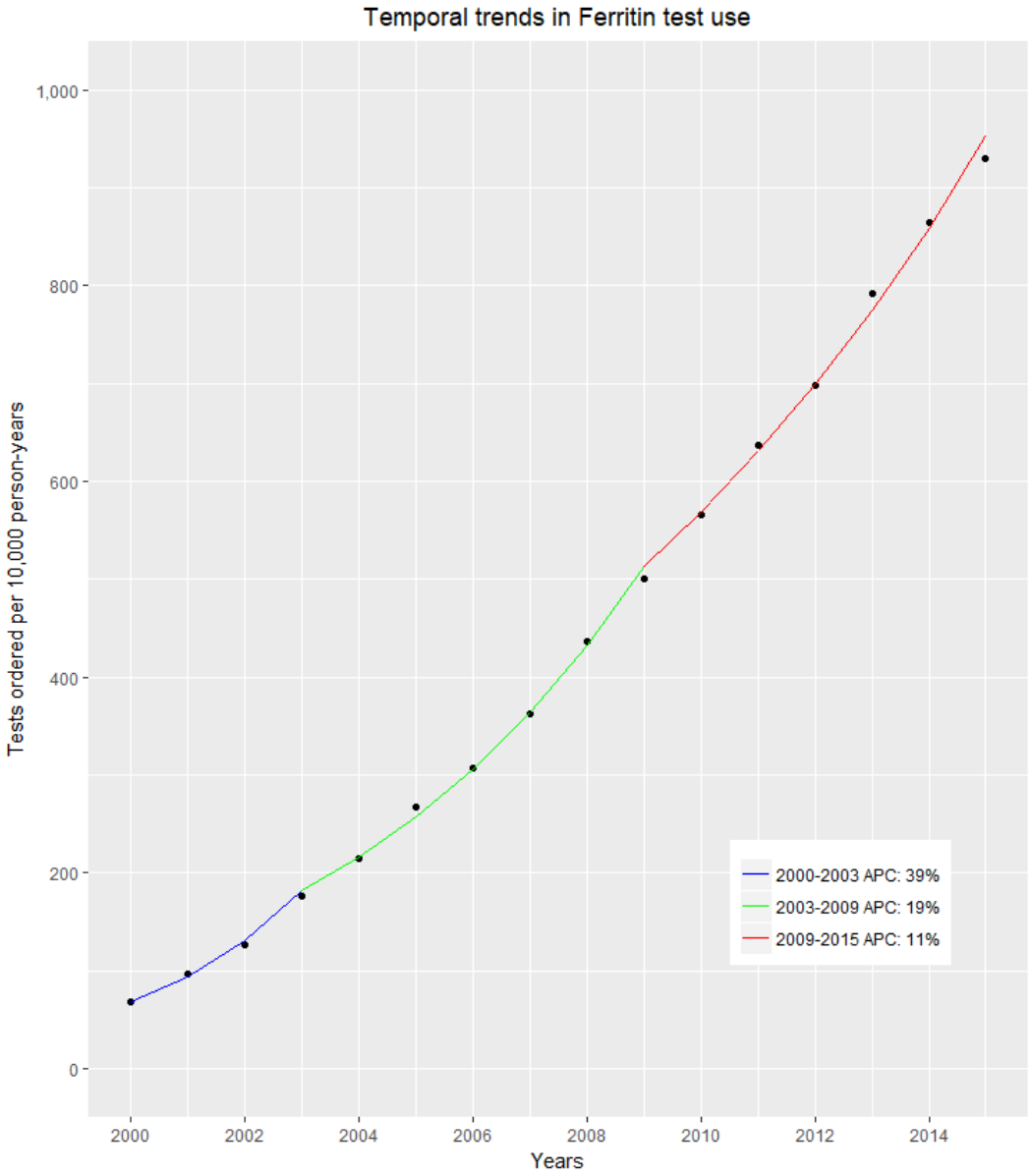

Temporal trends in Chest x-ray use
Tests ordered per 10,000 person-years
600
500
400
300
200
100
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2015 APC: 3.9%


Temporal trends in Vitamin B12 test use
Tests ordered per 10,000 person-years
800
600
400
200
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2007 APC: 26%
2007-2013 APC: 16%
2013-2015 APC: 9.5%

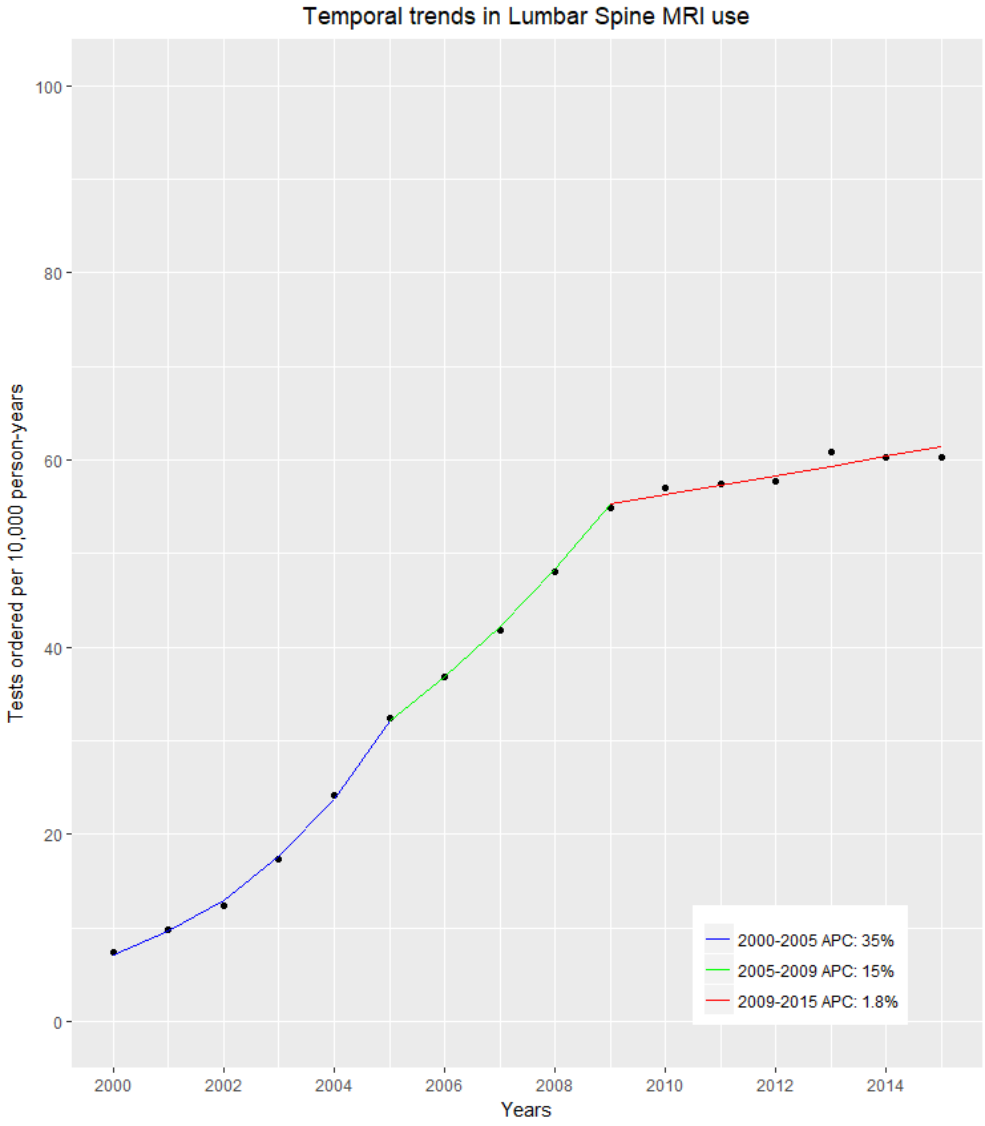

Temporal trends in CRP use
Tests ordered per 10,000 person-years
1,000
800
600
400
200
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 39%
2004-2008 APC: 14%
2008-2015 APC: 7.1%

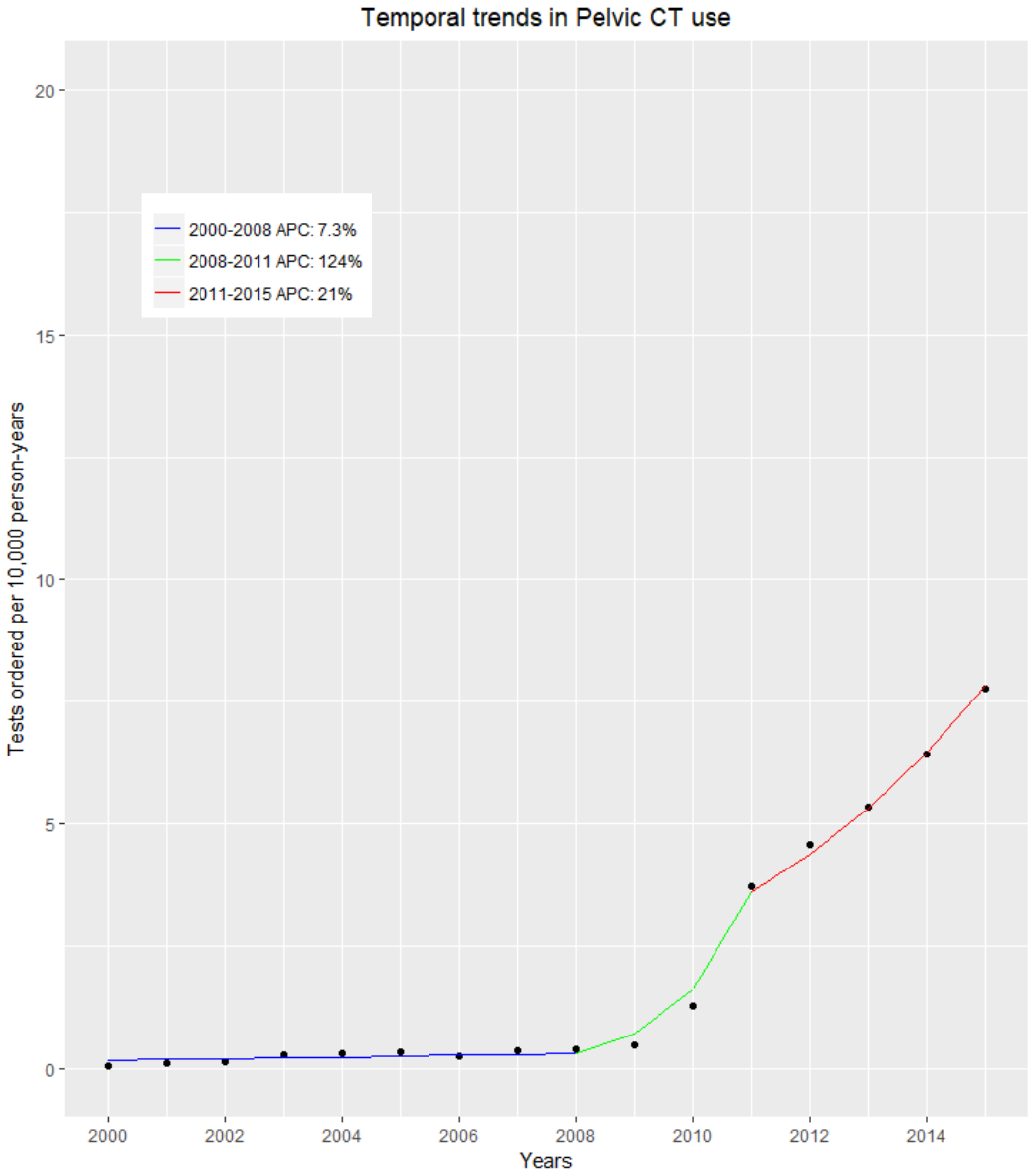

Temporal trends in Iron test use
Tests ordered per 10,000 person-years
200
150
100
50
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2008 APC: 21%
2008-2013 APC: 9.0%
2013-2015 APC: 17.0%

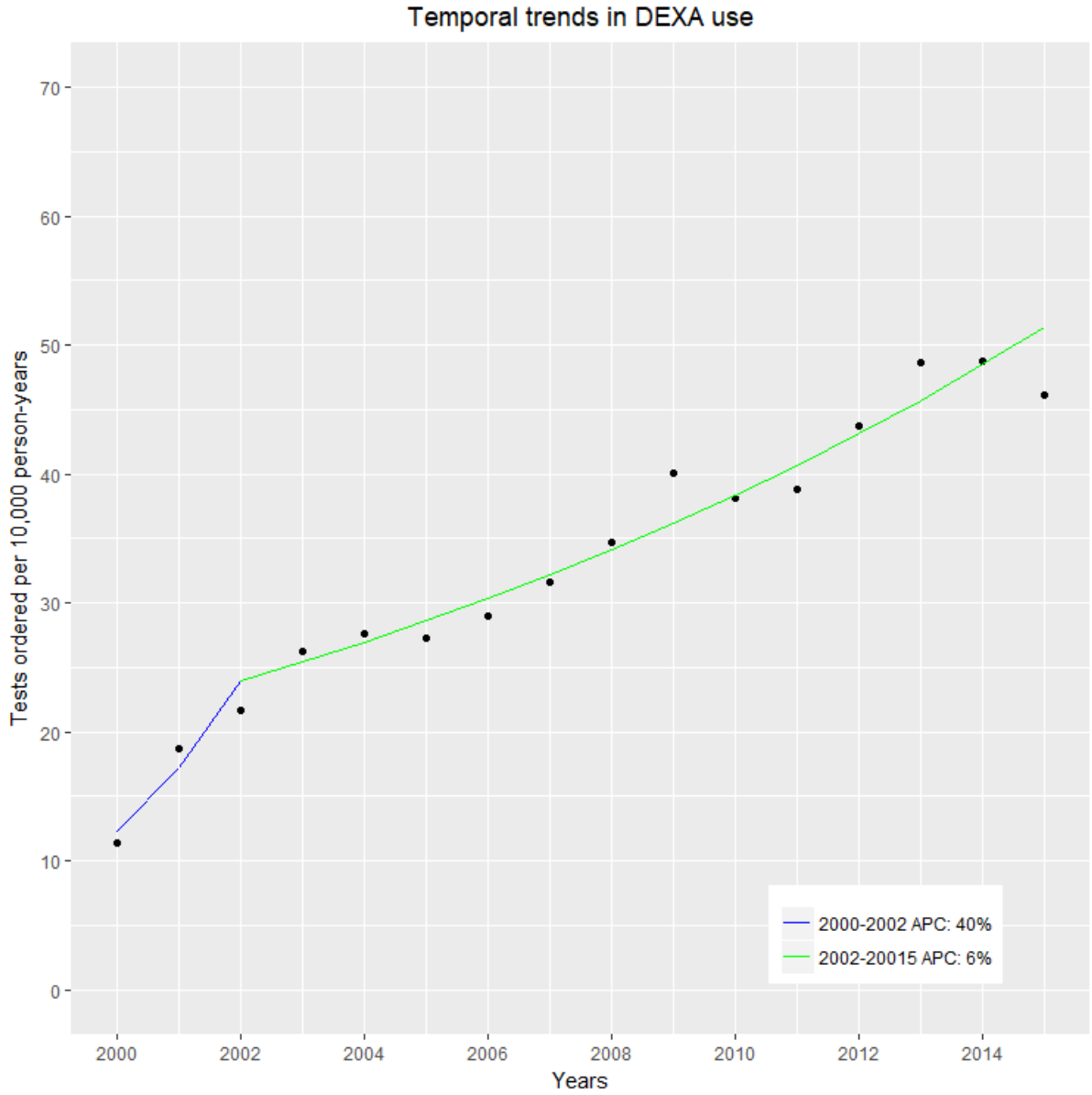

Temporal trends in Folate test use
Tests ordered per 10,000 person-years
1,000
800
600
400
200
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 28%
2004-2012 APC: 16%
2012-2015 APC: 9.4%

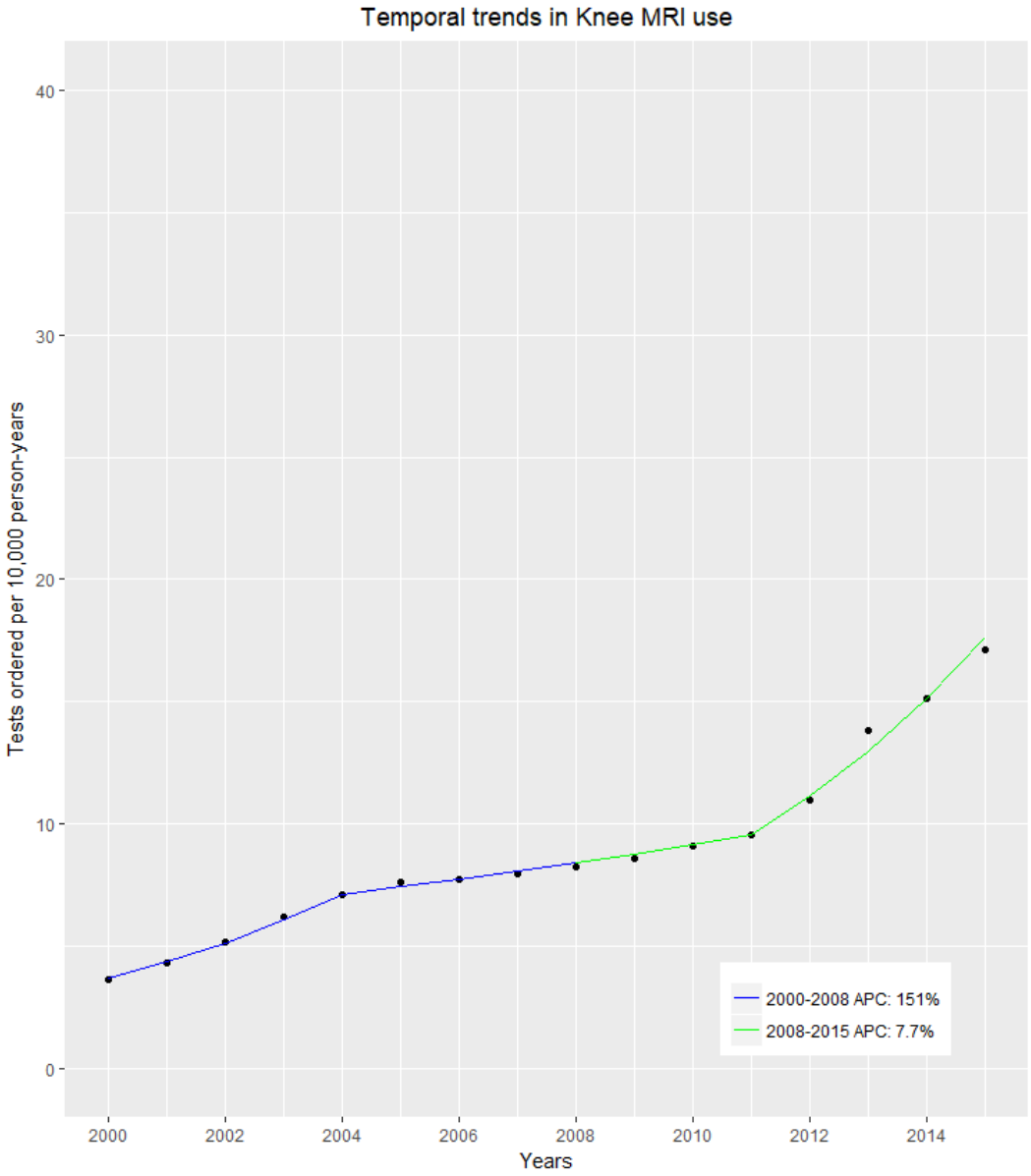


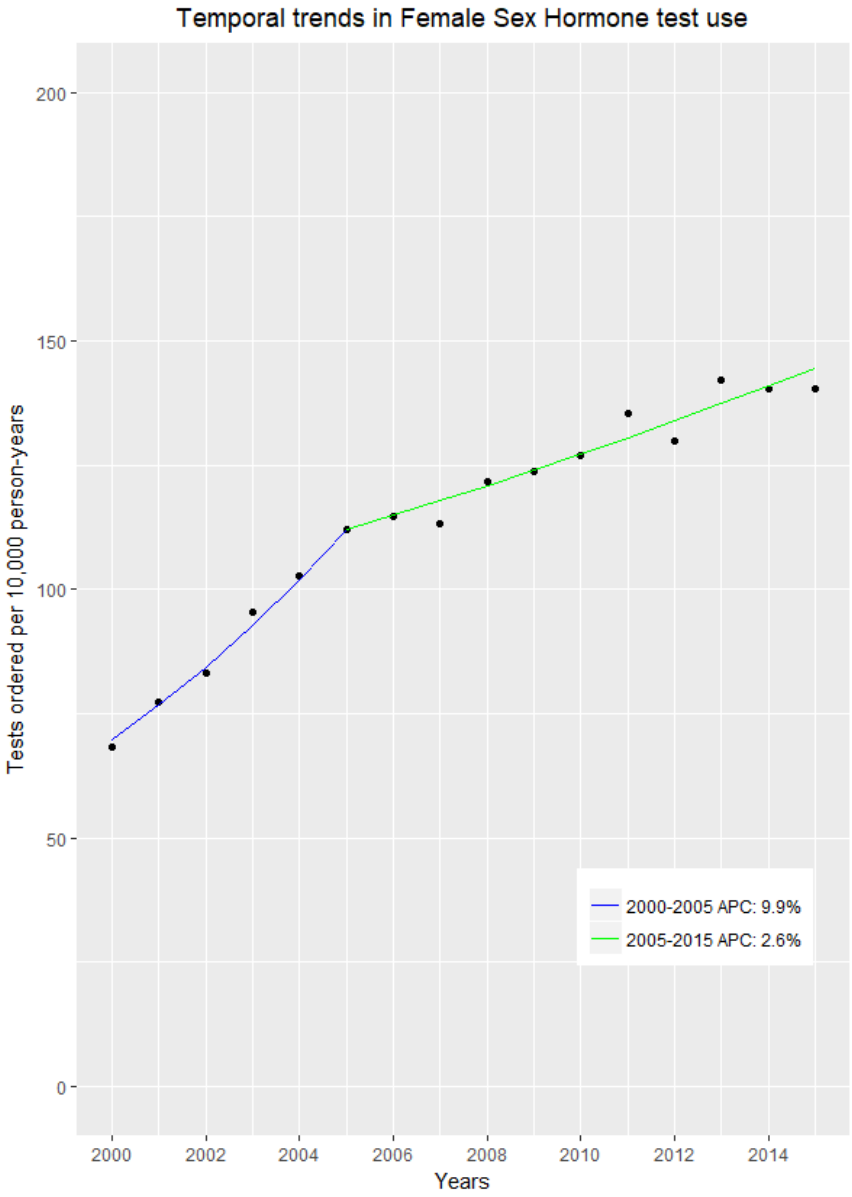
Pattern: Rapid increase and then increased less quickly

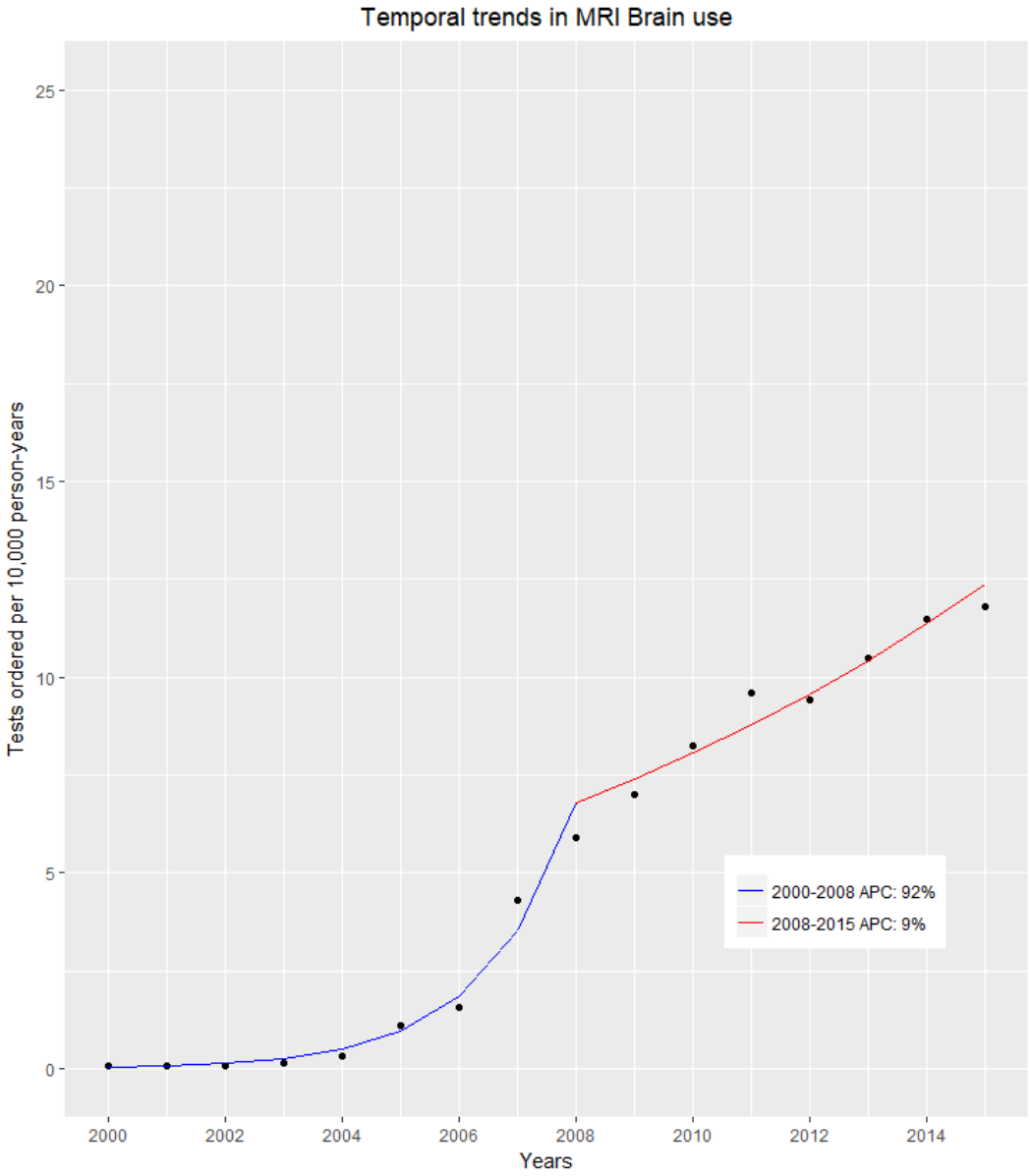

Temporal trends in Lumbar Spine MRI use
Tests ordered per 10,000 person-years
100
80
60
40
20
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2005 APC: 35%
2005-2009 APC: 15%
2009-2015 APC: 1.8%

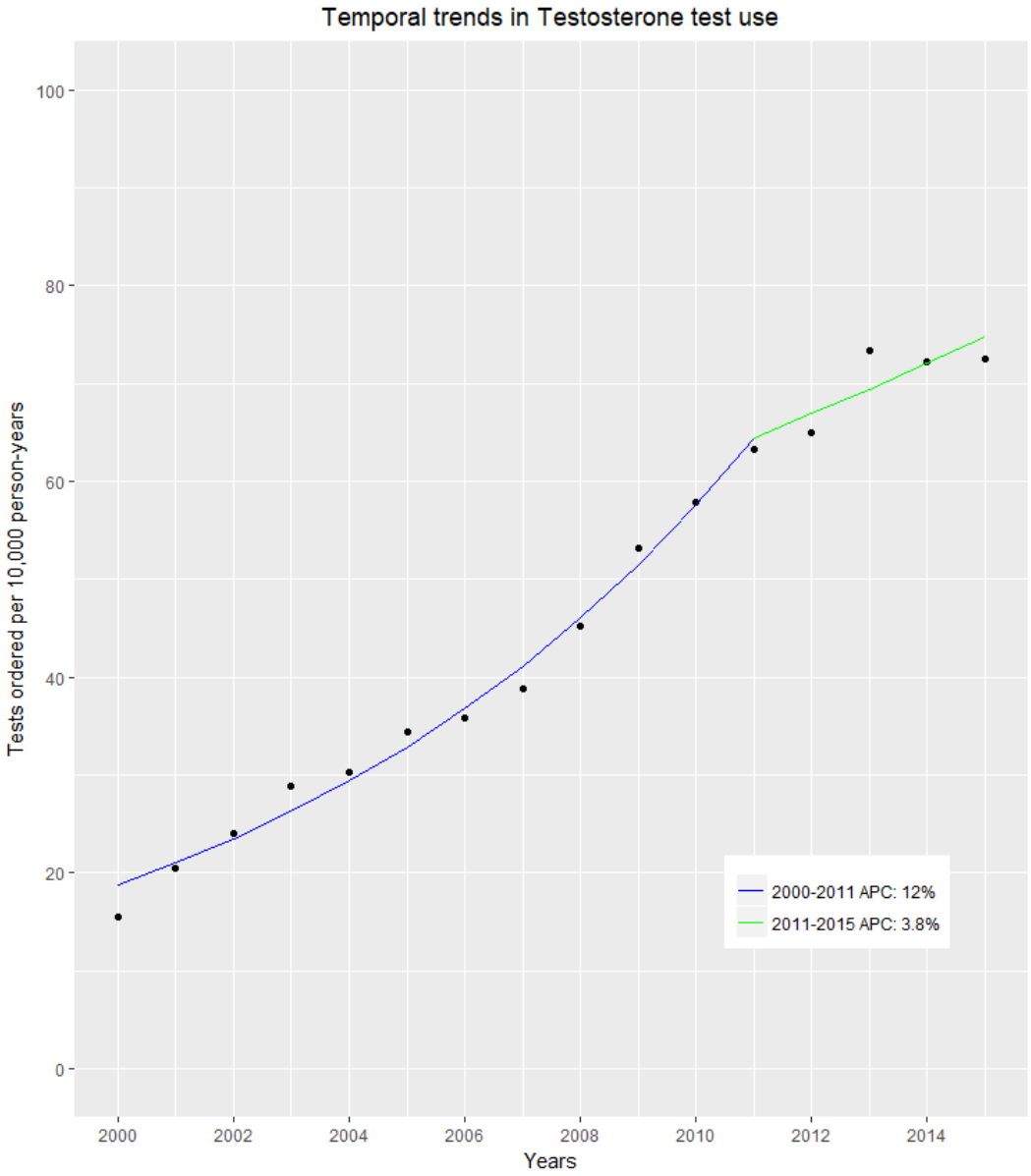

Temporal trends in Pelvic CT use
2000-2008 APC: 7.3%
2008-2011 APC: 124%
2011-2015 APC: 21%
Tests ordered per 10,000 person-years
Years
0
5
10
15
20
2000
2002
2004
2006
2008
2010
2012
2014

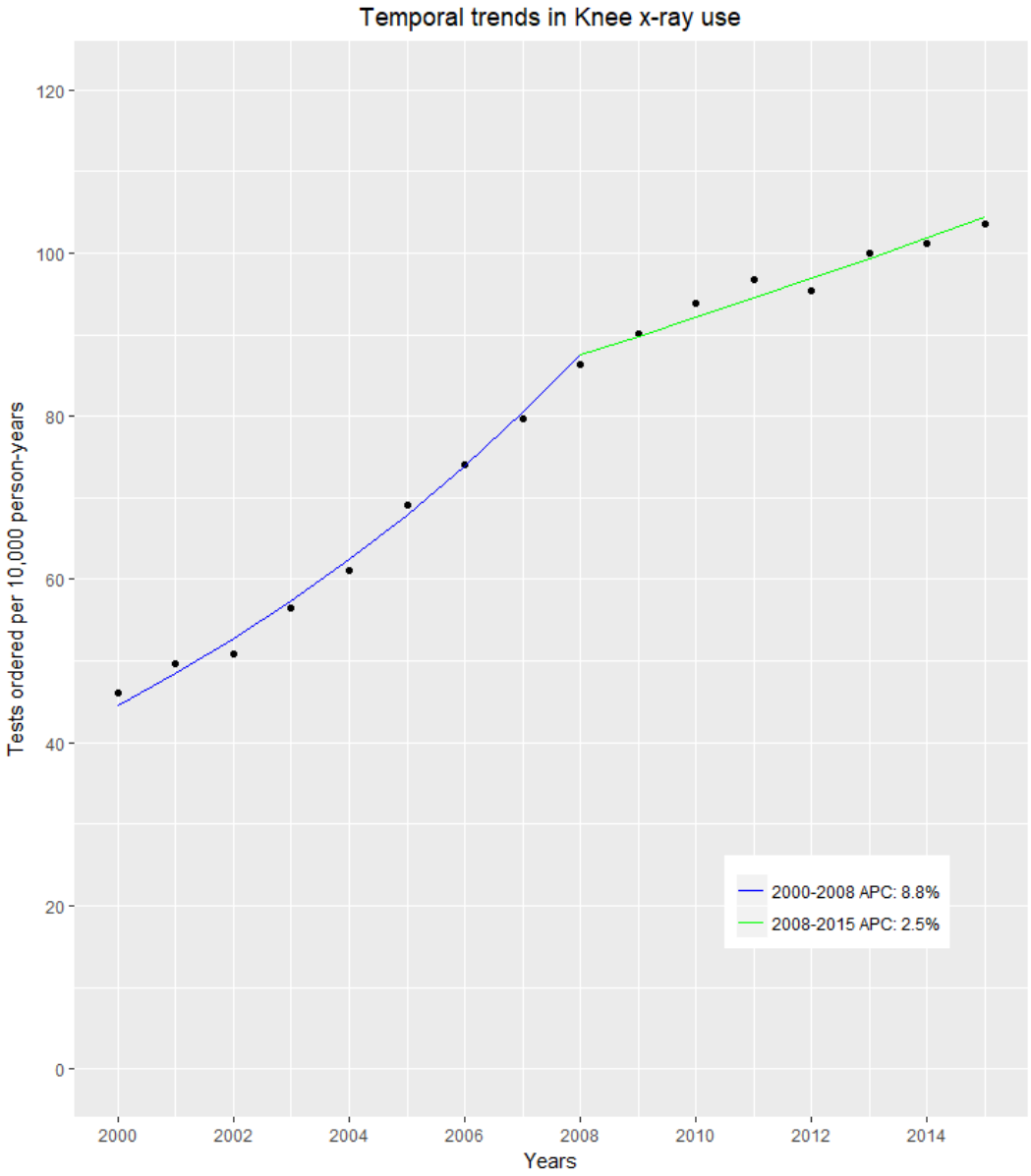

Temporal trends in DEXA use
Tests ordered per 10,000 person-years
70
60
50
40
30
20
10
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2002 APC: 40%
2002-20015 APC: 6%

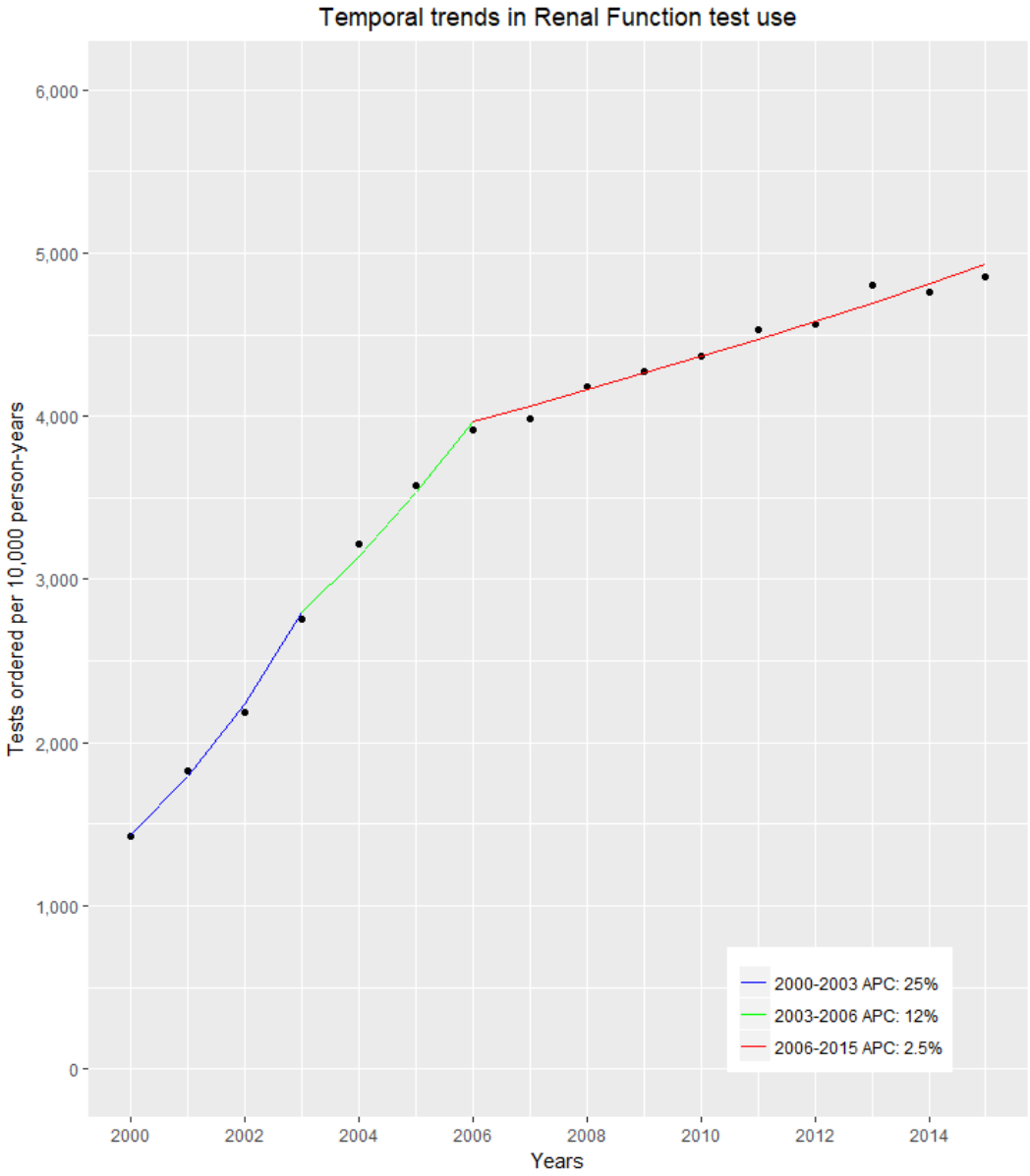

Temporal trends in Knee MRI use
Tests ordered per 10,000 person-years
40
30
20
10
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2008 APC: 151%
2008-2015 APC: 7.7%

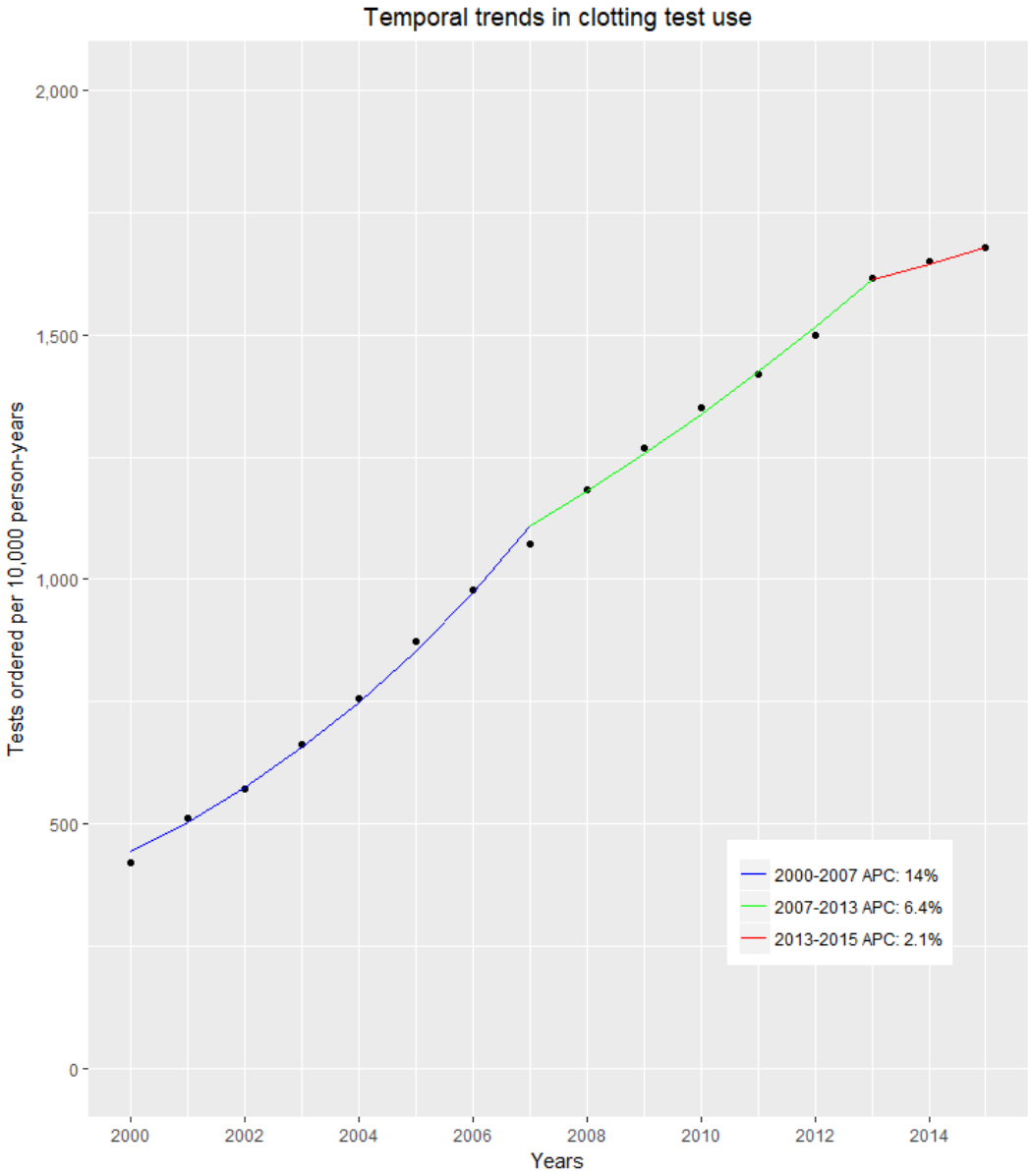

Temporal trends in Female Sex Hormone test use
Tests ordered per 10,000 person-years
200
150
100
50
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2005 APC: 9.9%
2005-2015 APC: 2.6%

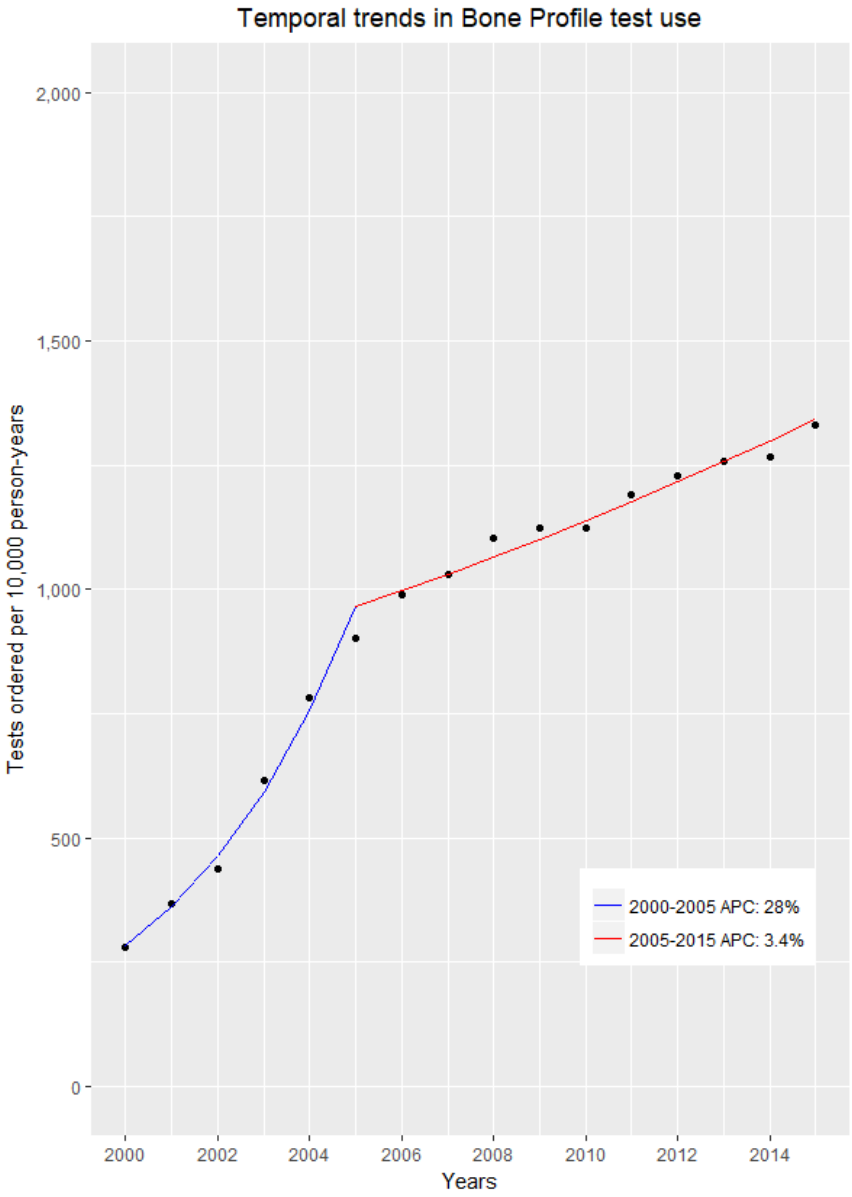

Temporal trends in MRI Brain use
Tests ordered per 10,000 person-years
25
20
15
10
5
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2008 APC: 92%
2008-2015 APC: 9%

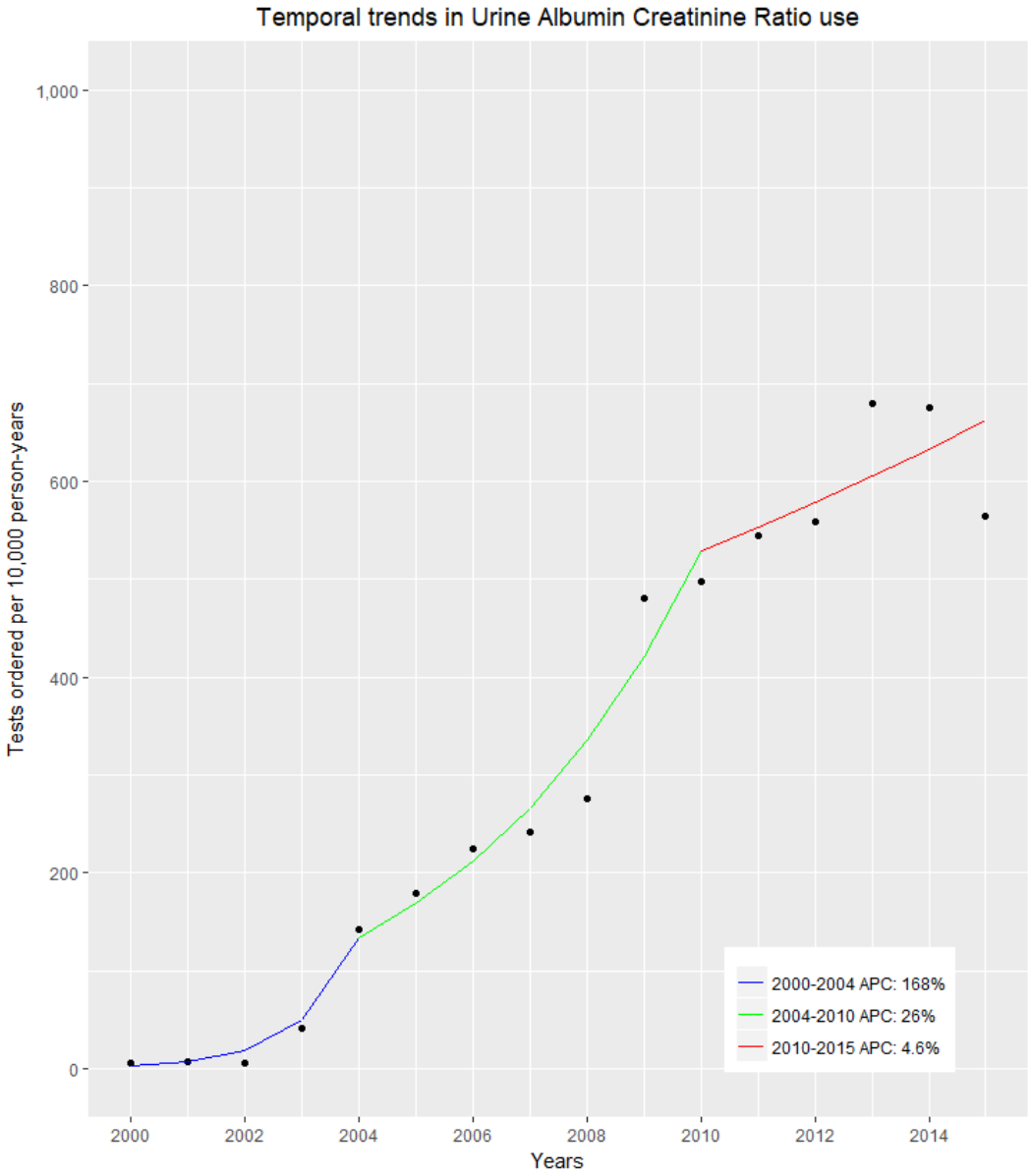

Temporal trends in Testosterone test use
Tests ordered per 10,000 person-years
100
80
60
40
20
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2011 APC: 12%
2011-2015 APC: 3.8%

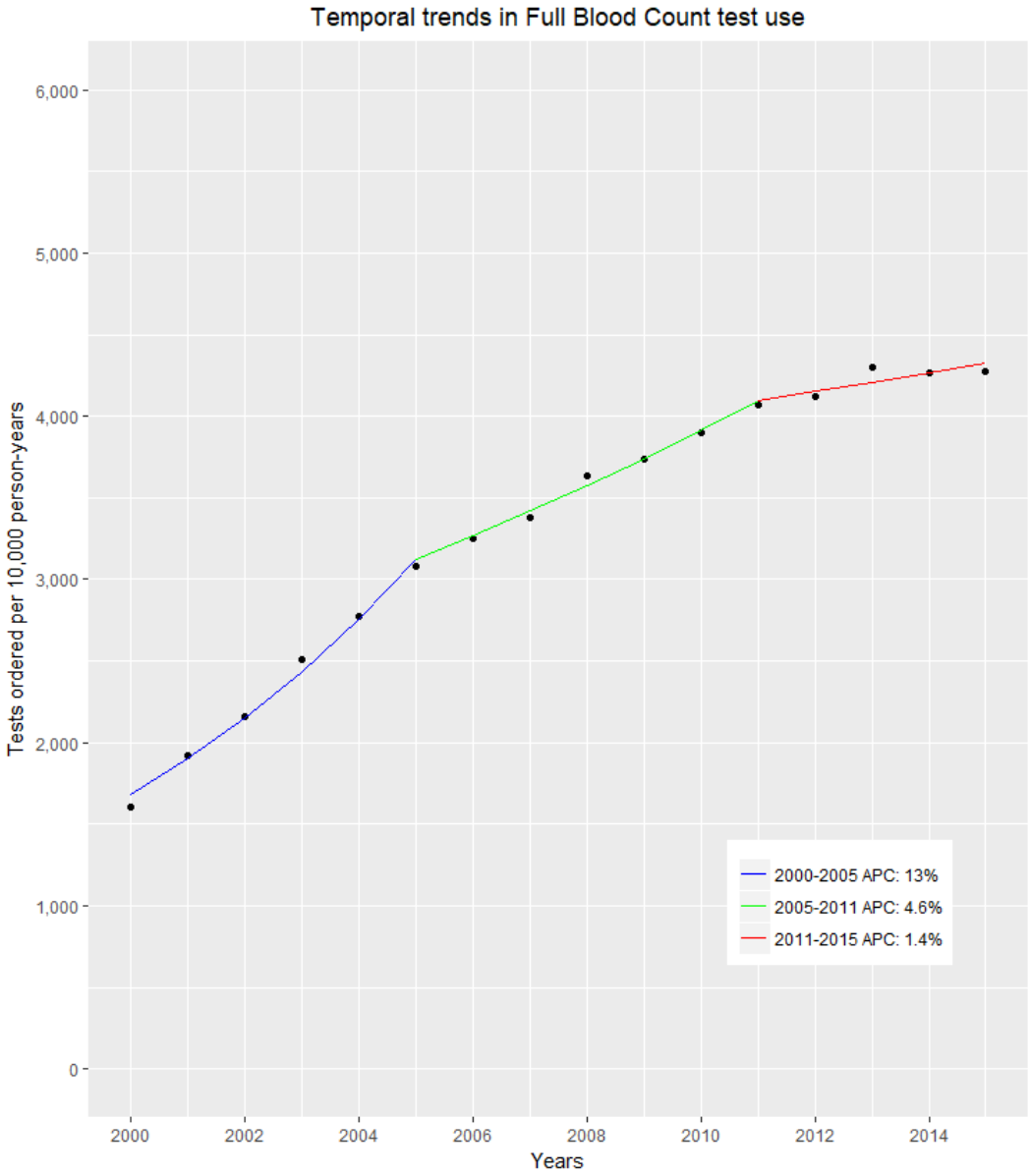

Temporal trends in Knee x-ray use
Tests ordered per 10,000 person-years
120
100
80
60
40
20
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2008 APC: 8.8%
2008-2015 APC: 2.5%

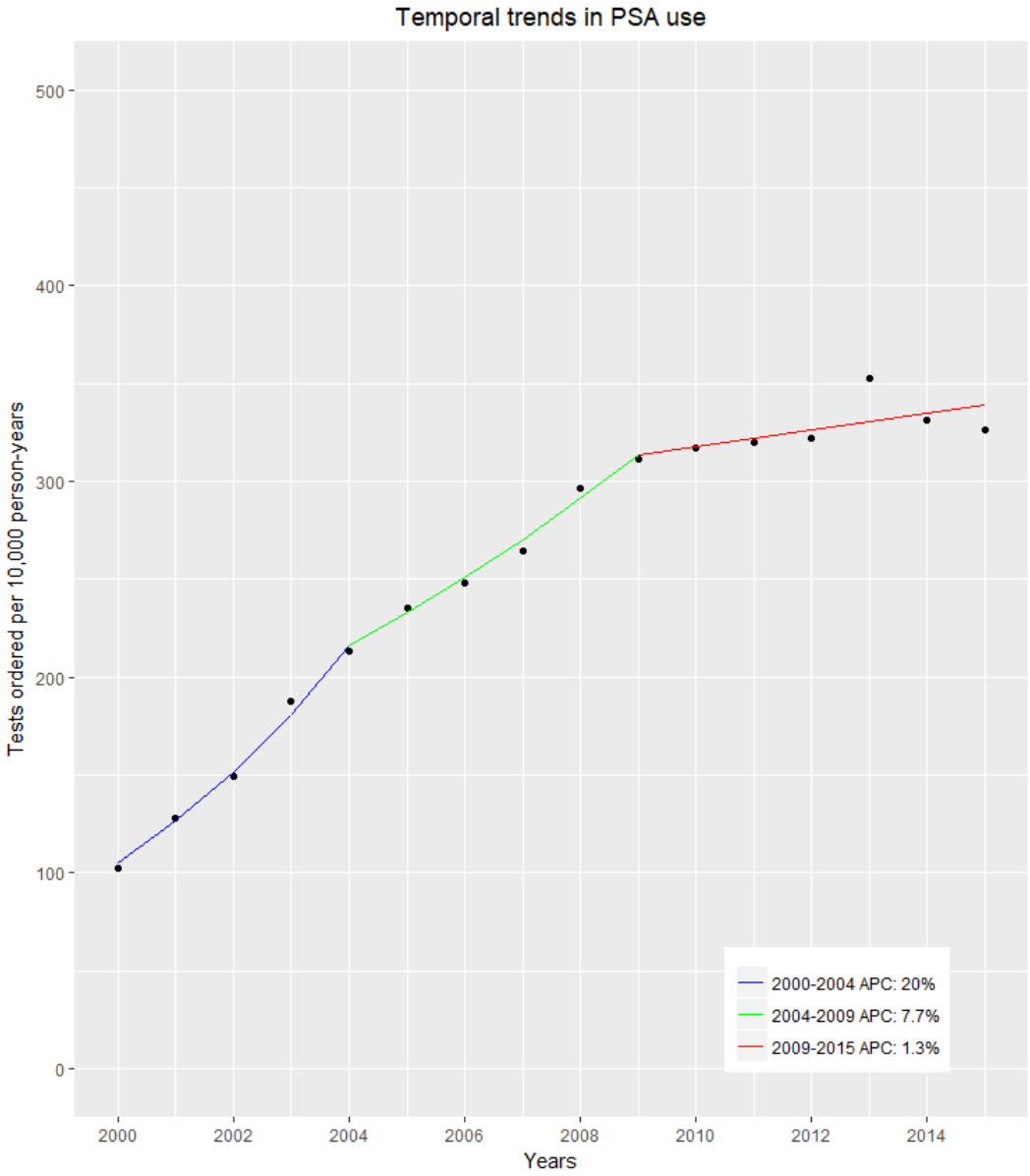

Temporal trends in Renal Function test use
Tests ordered per 10,000 person-years
6,000
5,000
4,000
3,000
2,000
1,000
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2003 APC: 25%
2003-2006 APC: 12%
2006-2015 APC: 2.5%

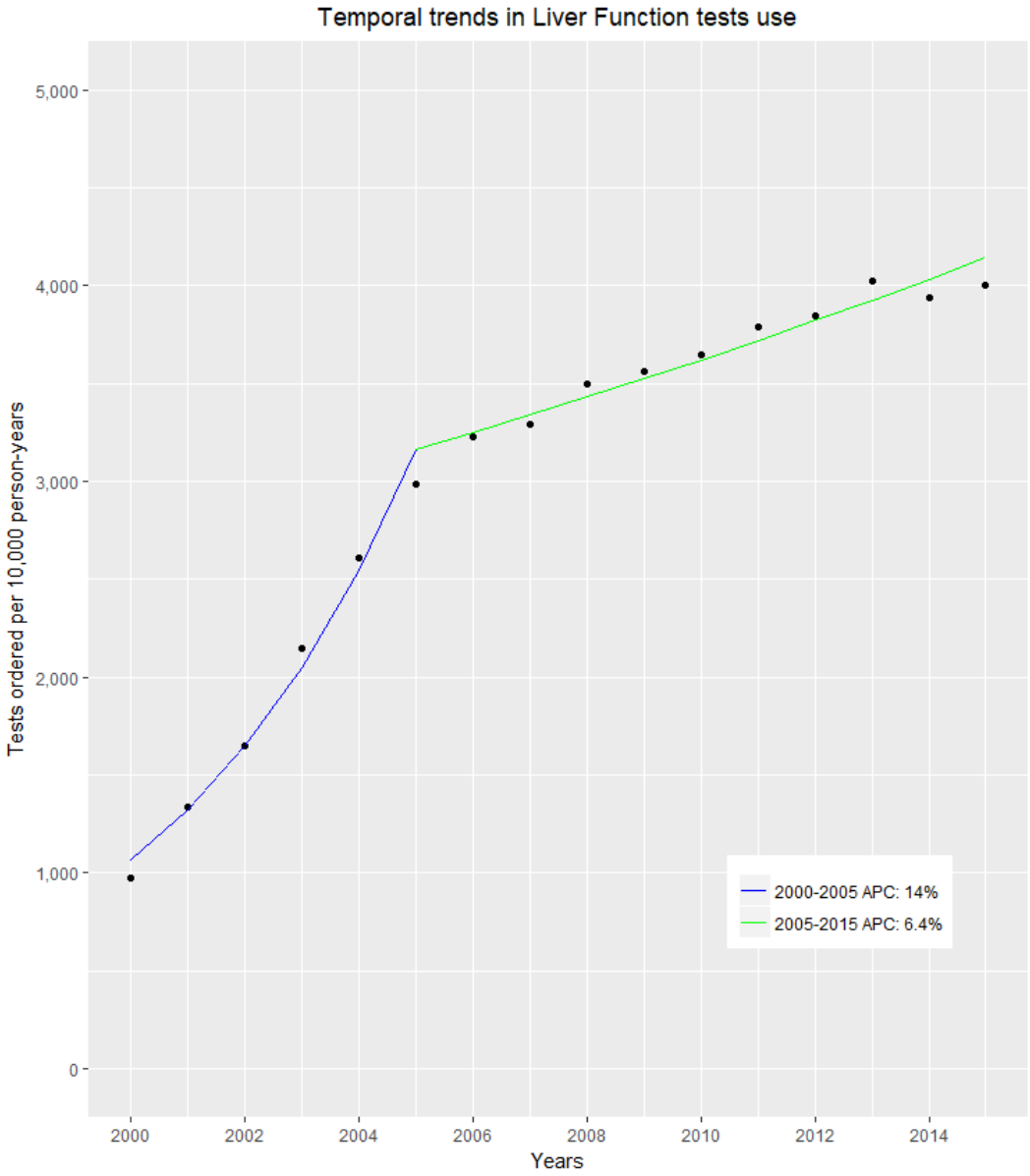

Temporal trends in clotting test use
Tests ordered per 10,000 person-years
2,000
1,500
1,000
500
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2007 APC: 14%
2007-2013 APC: 6.4%
2013-2015 APC: 2.1%


Temporal trends in Bone Profile test use
Tests ordered per 10,000 person-years
2,000
1,500
1,000
500
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2005 APC: 28%
2005-2015 APC: 3.4%

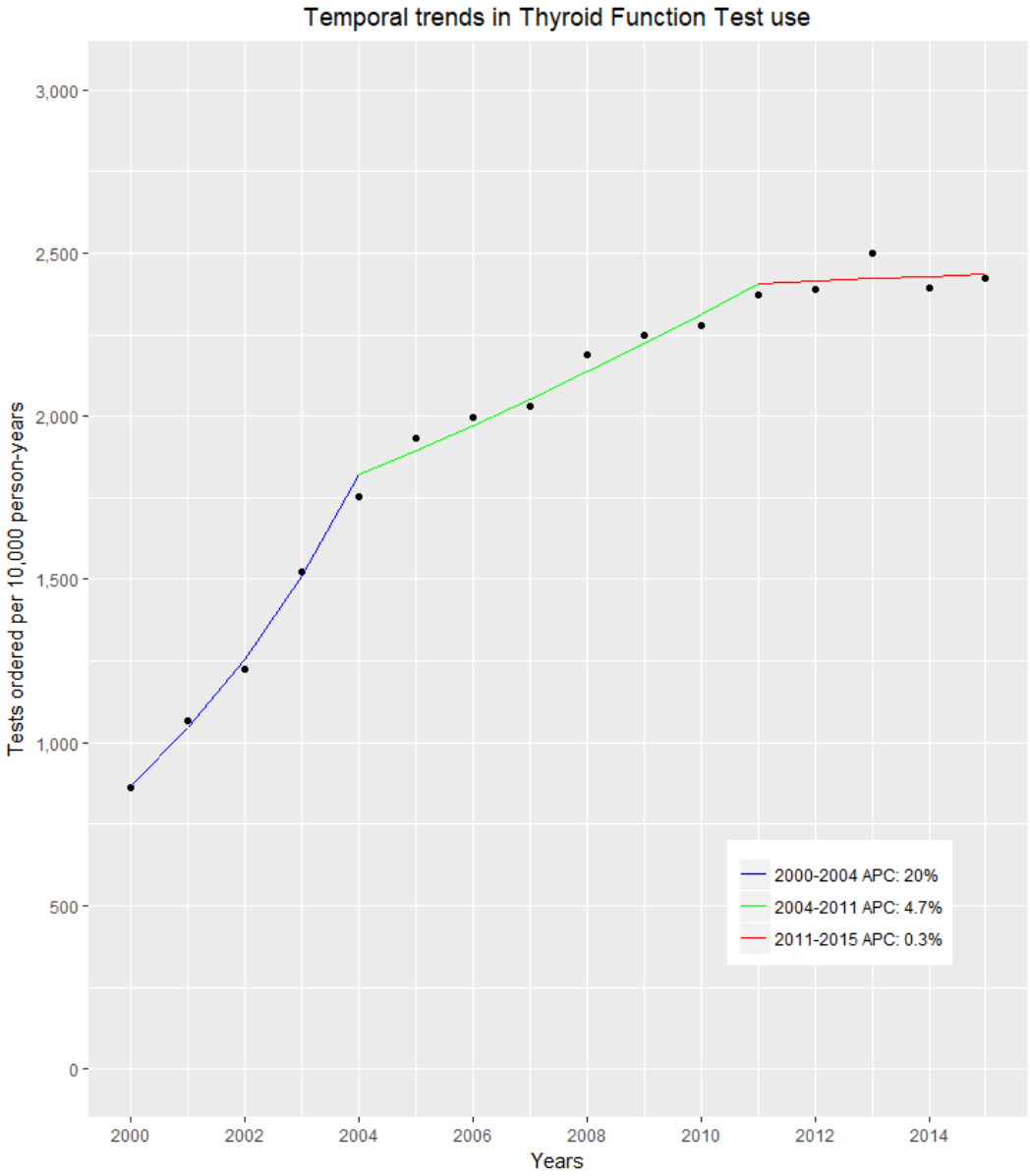

Temporal trends in Urine Albumin Creatinine Ratio use
Tests ordered per 10,000 person-years
1,000
800
600
400
200
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 168%
2004-2010 APC: 26%
2010-2015 APC: 4.6%

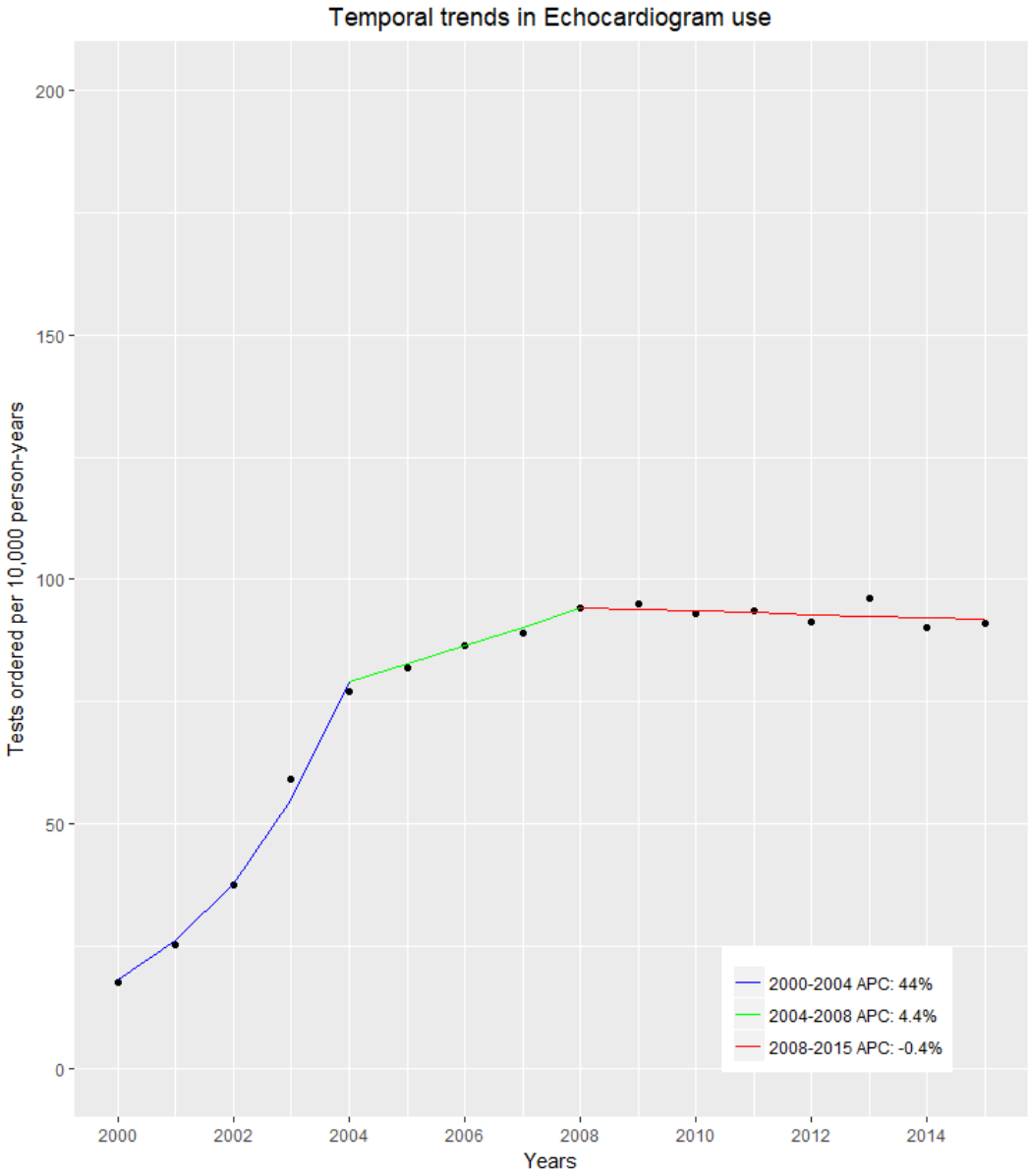

Temporal trends in Full Blood Count test use
Tests ordered per 10,000 person-years
6,000
5,000
4,000
3,000
2,000
1,000
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2005 APC: 13%
2005-2011 APC: 4.6%
2011-2015 APC: 1.4%

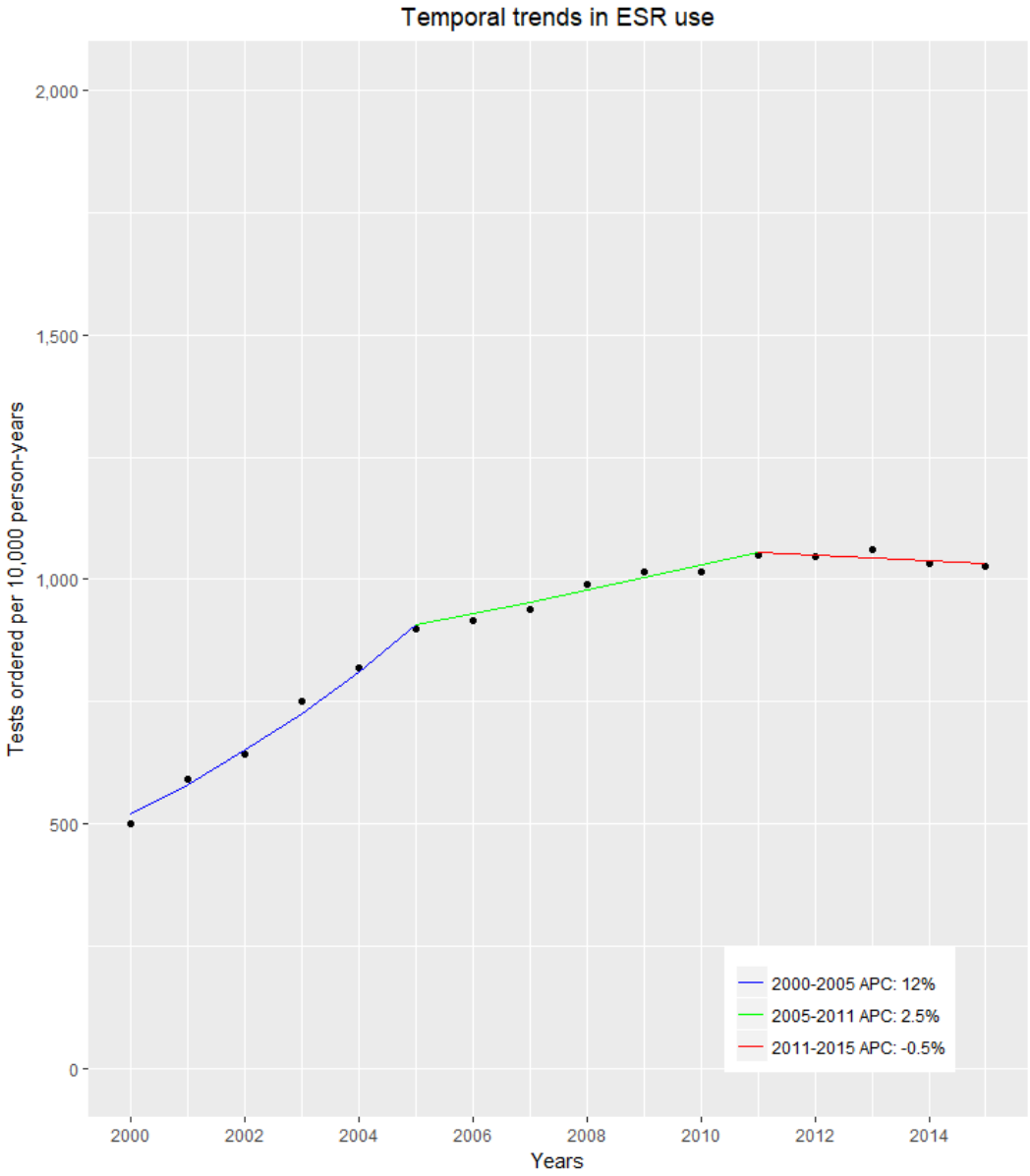

Temporal trends in PSA use
Tests ordered per 10,000 person-years
500
400
300
200
100
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 20%
2004-2009 APC: 7.7%
2009-2015 APC: 1.3%

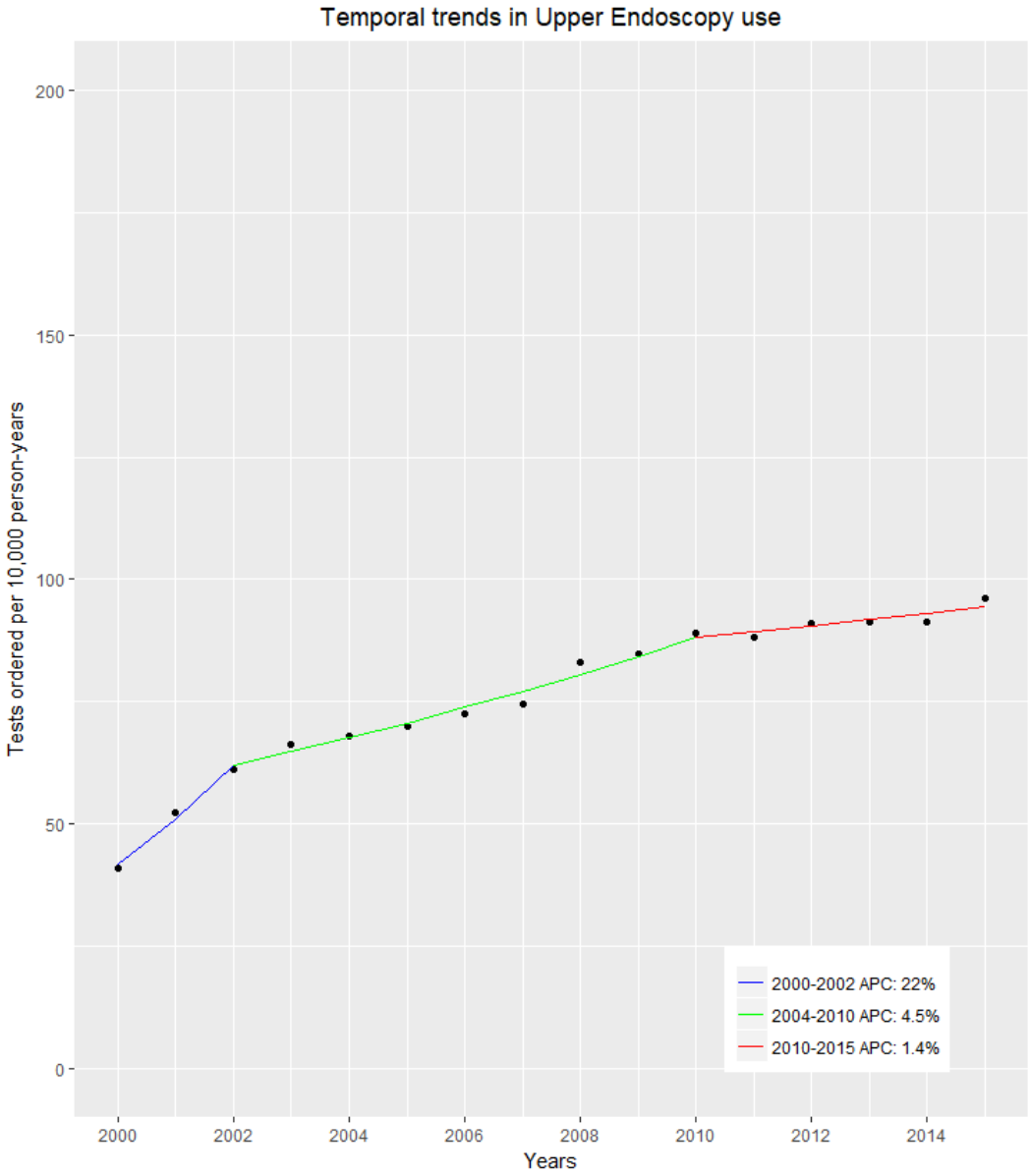


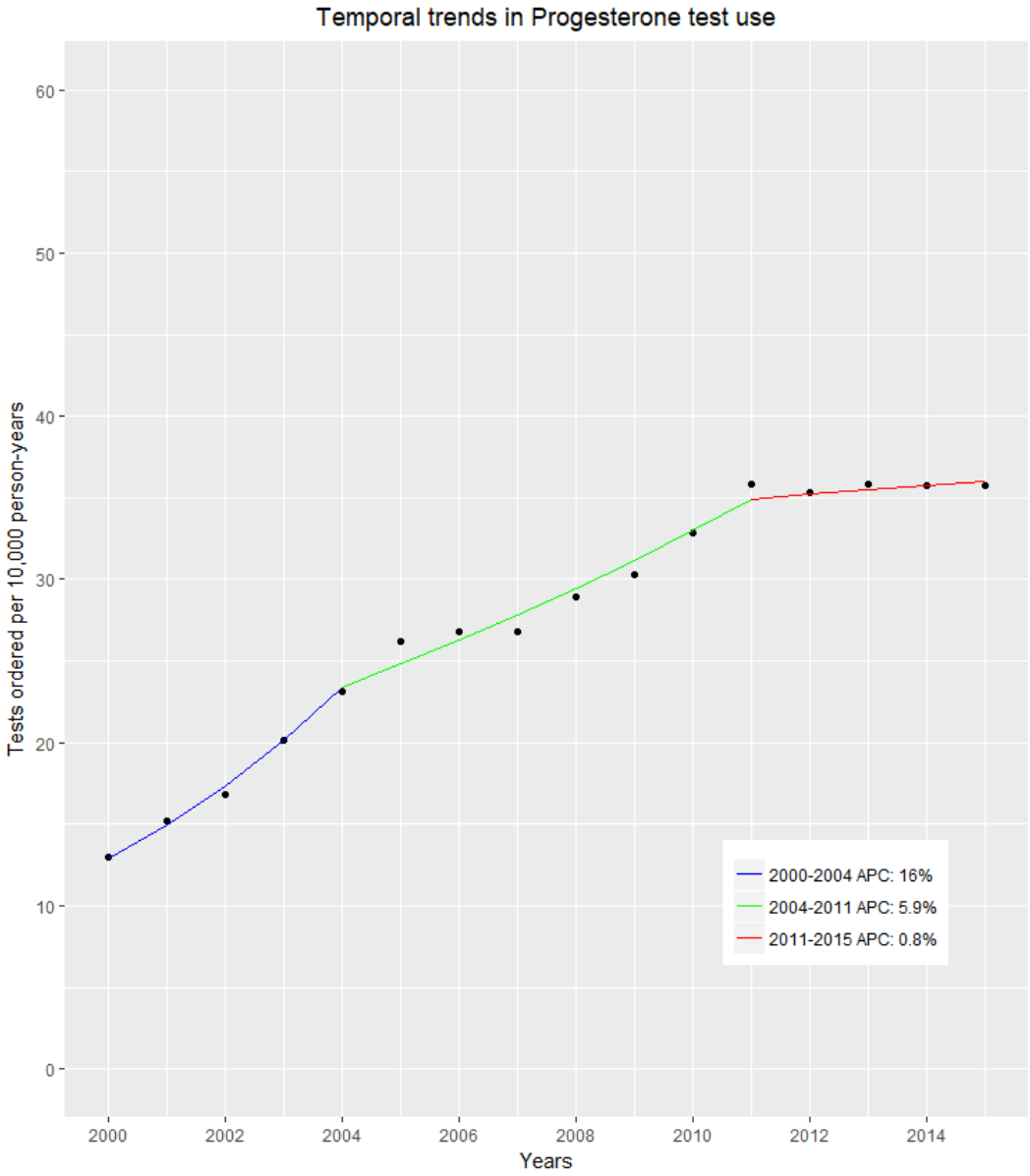
Pattern: Plateau

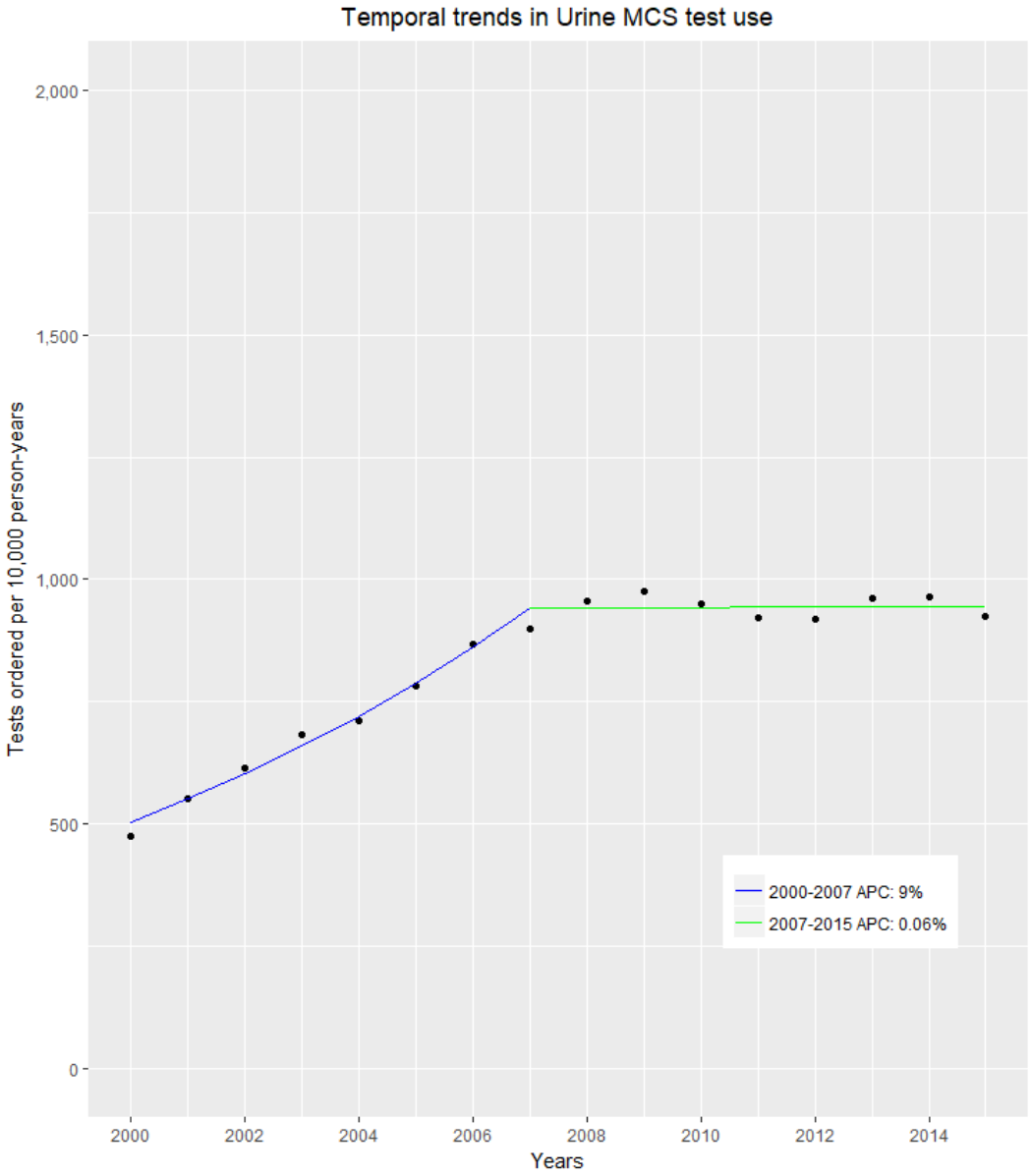

Temporal trends in Thyroid Function Test use
Tests ordered per 10,000 person-years
3,000
2,500
2,000
1,500
1,000
500
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 20%
2004-2011 APC: 4.7%
2011-2015 APC: 0.3%

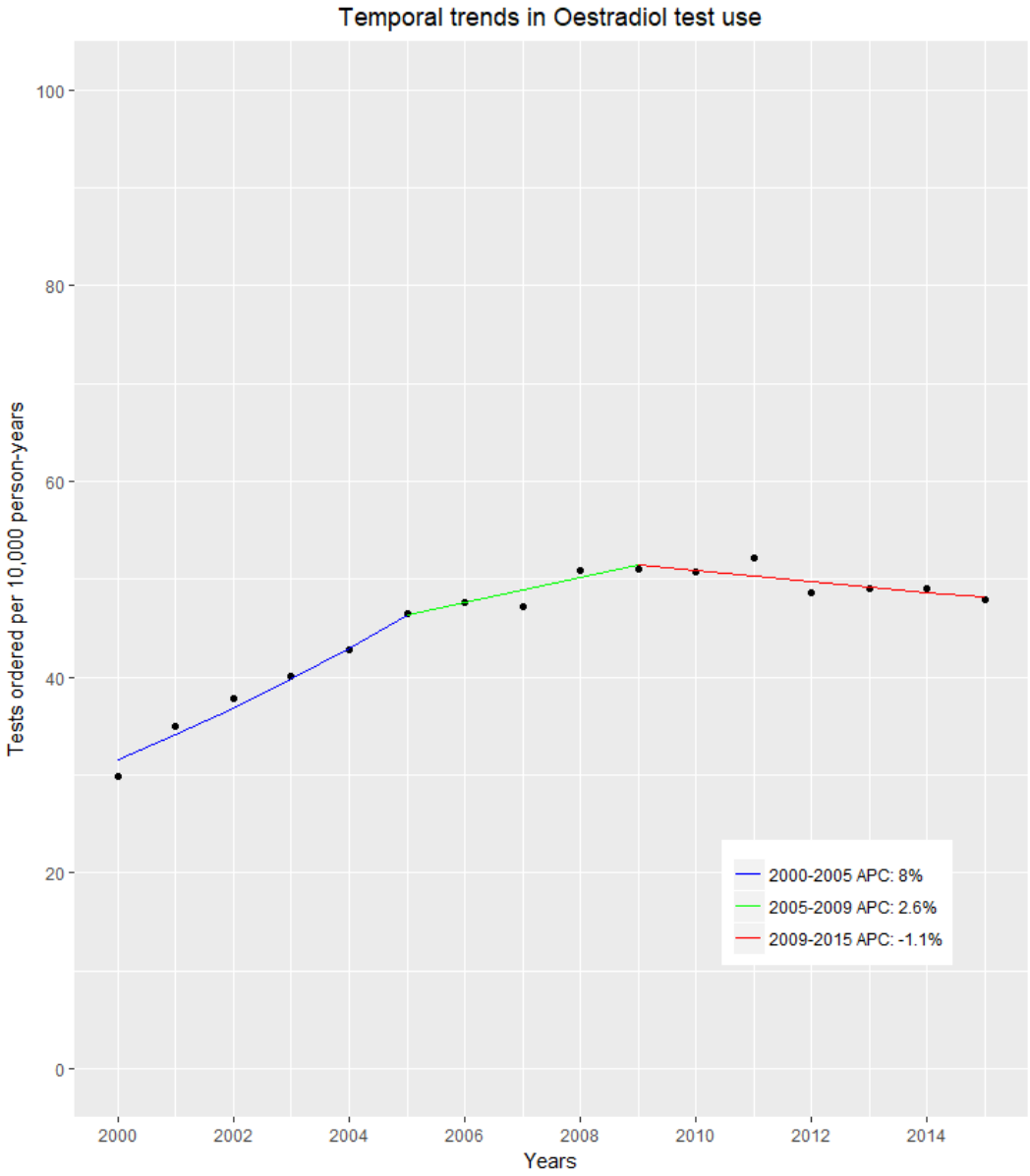

Temporal trends in Echocardiogram use
Tests ordered per 10,000 person-years
200
150
100
50
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 44%
2004-2008 APC: 4.4%
2008-2015 APC: -0.4%

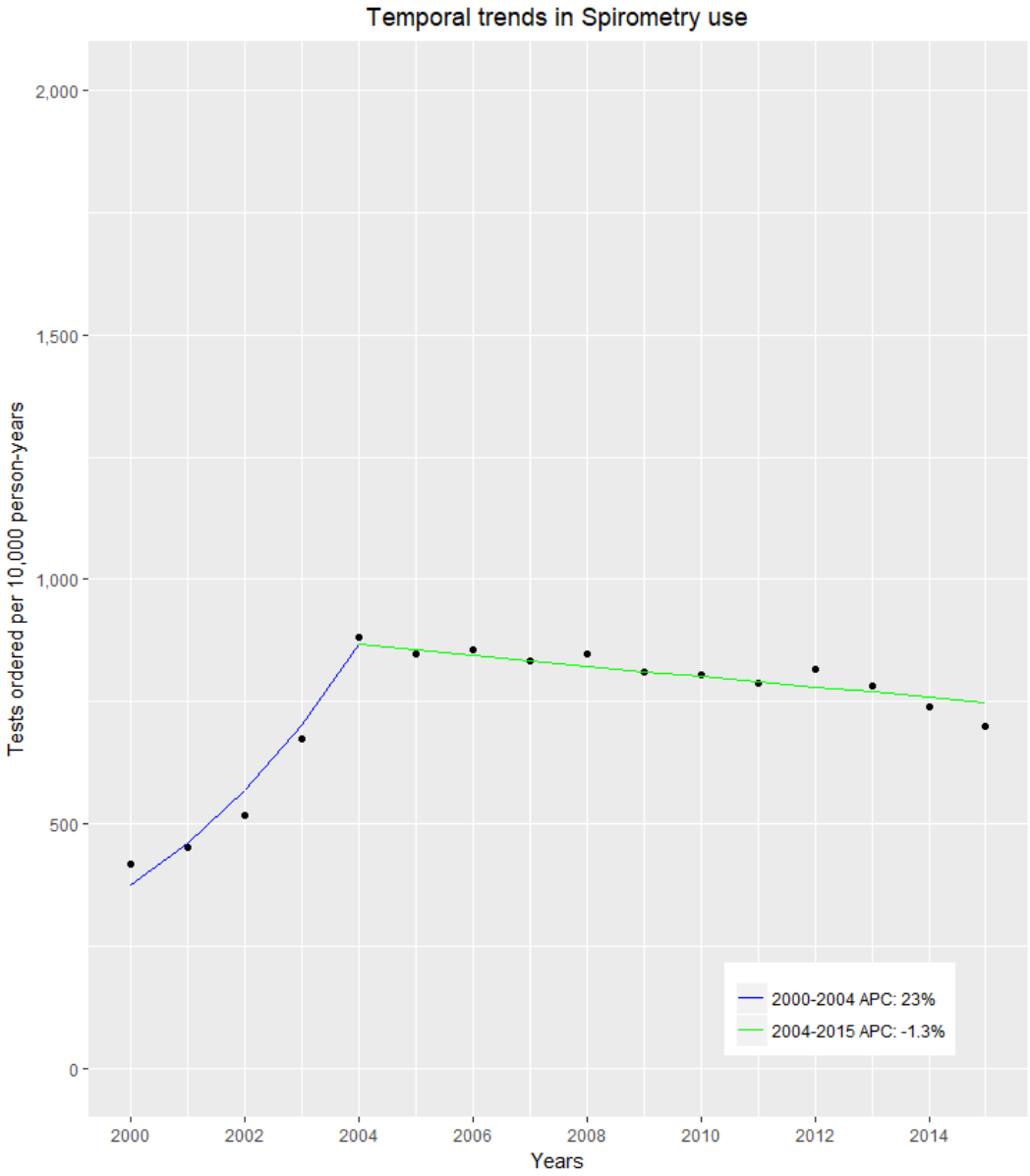

Temporal trends in ESR use
Tests ordered per 10,000 person-years
2,000
1,500
1,000
500
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2005 APC: 12%
2005-2011 APC: 2.5%
2011-2015 APC: -0.5%


Temporal trends in Upper Endoscopy use
Tests ordered per 10,000 person-years
200
150
100
50
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2002 APC: 22%
2004-2010 APC: 4.5%
2010-2015 APC: 1.4%

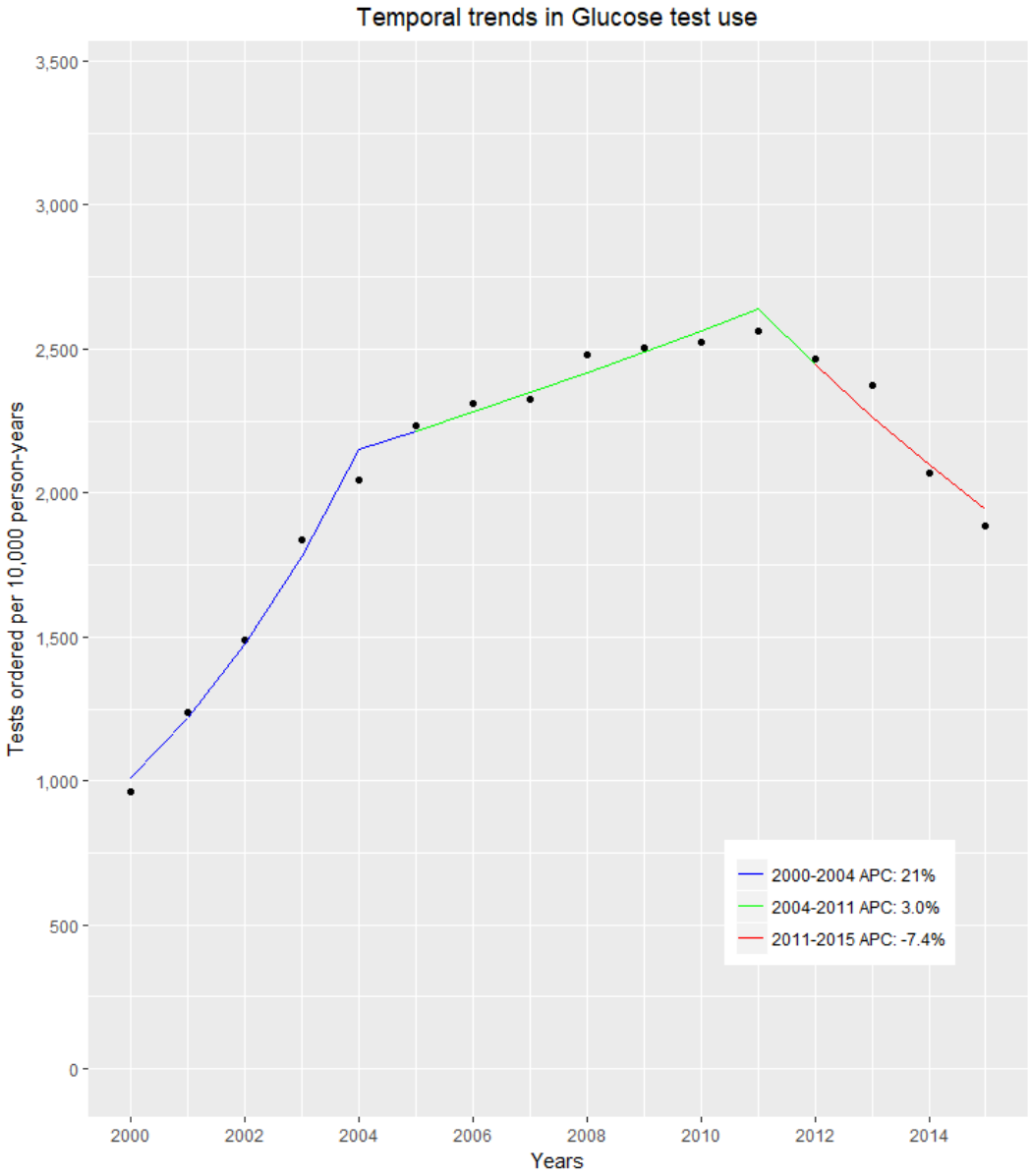

Temporal trends in Progesterone test use
Tests ordered per 10,000 person-years
Years
0
10
20
30
40
50
60
2000
2002
2004
2006
2008
2010
2012
2014
2000-2004 APC: 16%
2004-2011 APC: 5.9%
2011-2015 APC: 0.8%

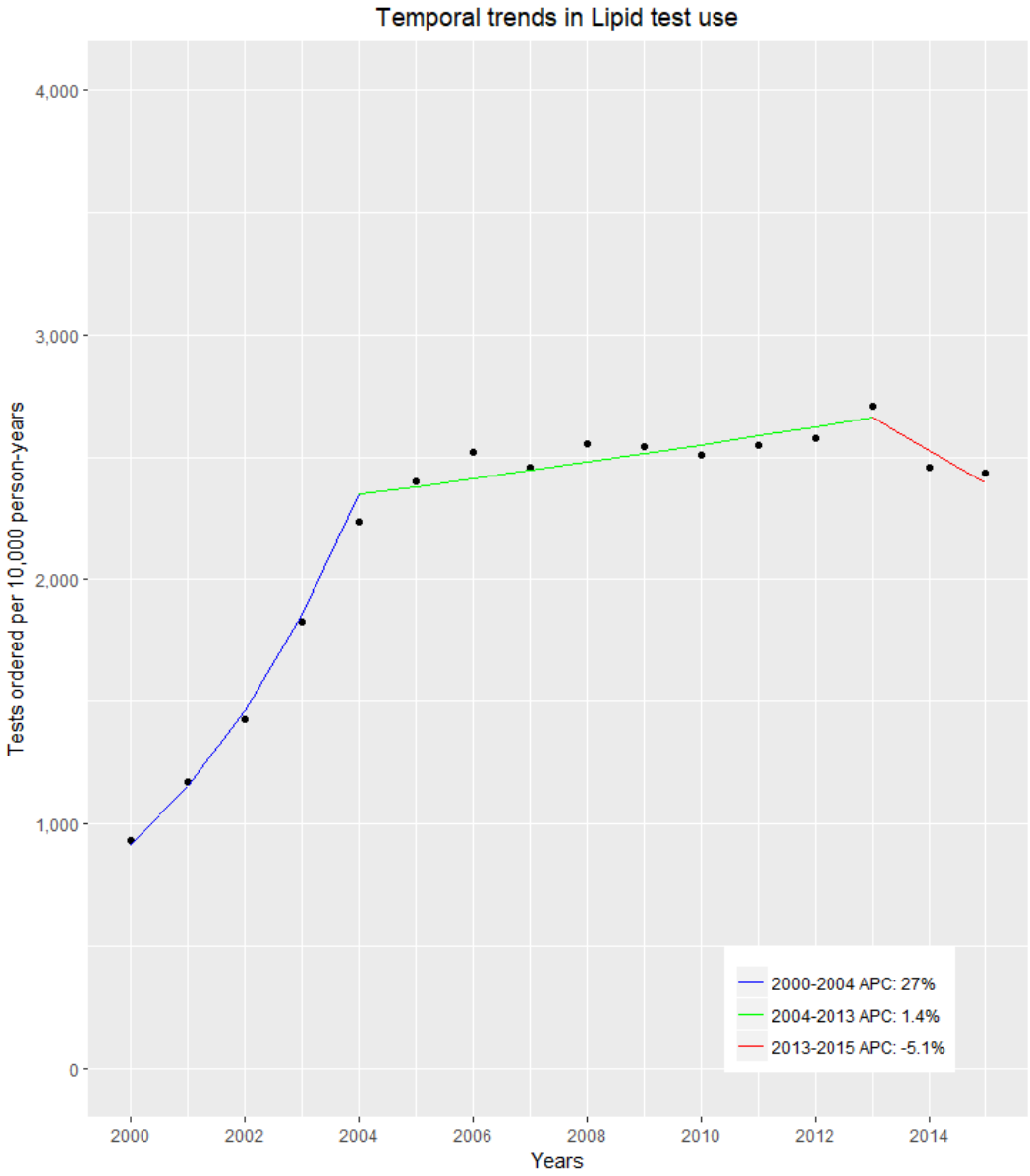

Temporal trends in Urine MCS test use
Tests ordered per 10,000 person-years
2,000
1,500
1,000
500
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2007 APC: 9%
2007-2015 APC: 0.06%

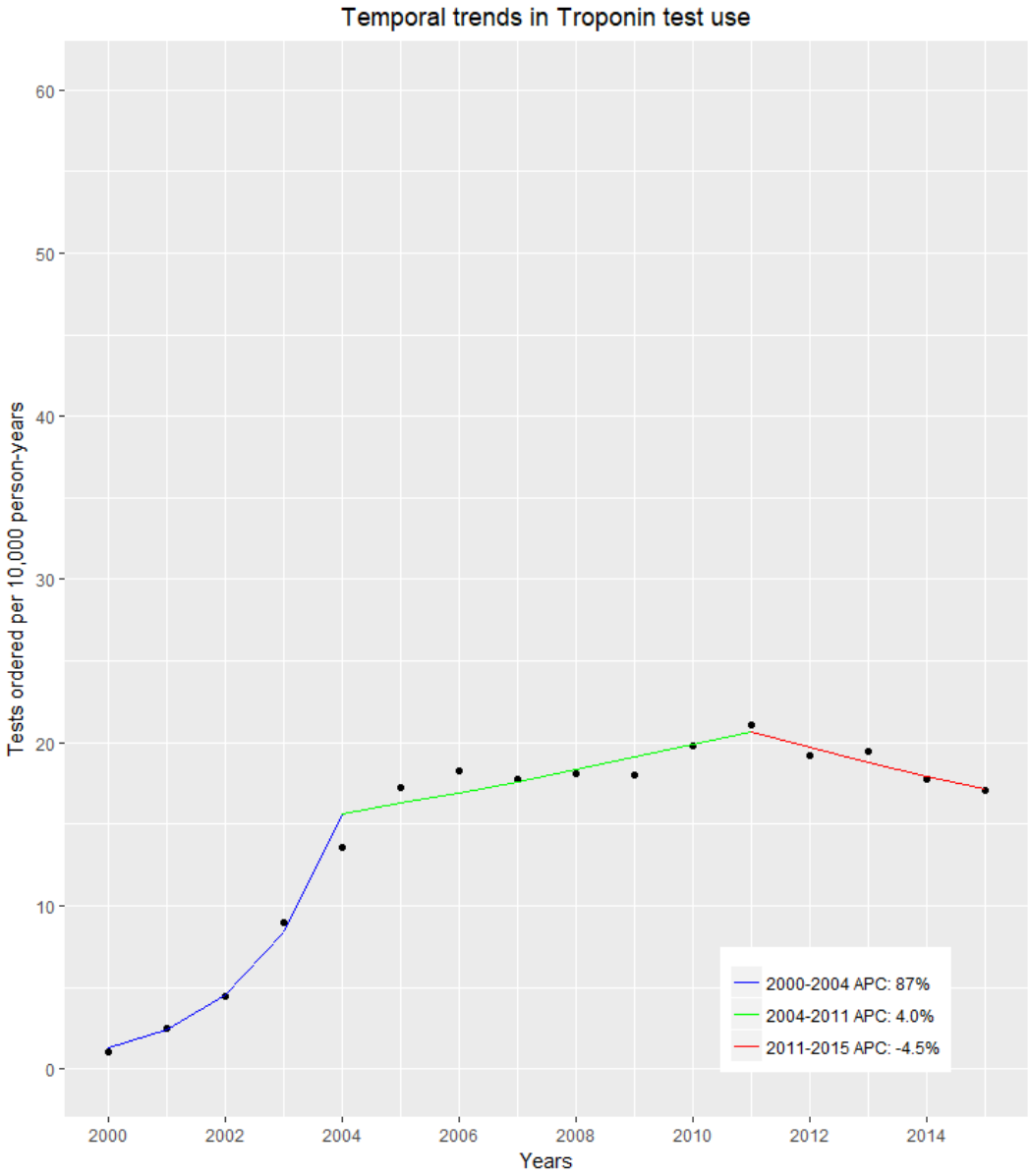

Temporal trends in Oestradiol test use
Tests ordered per 10,000 person-years
100
80
60
40
20
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2005 APC: 8%
2005-2009 APC: 2.6%
2009-2015 APC: -1.1%

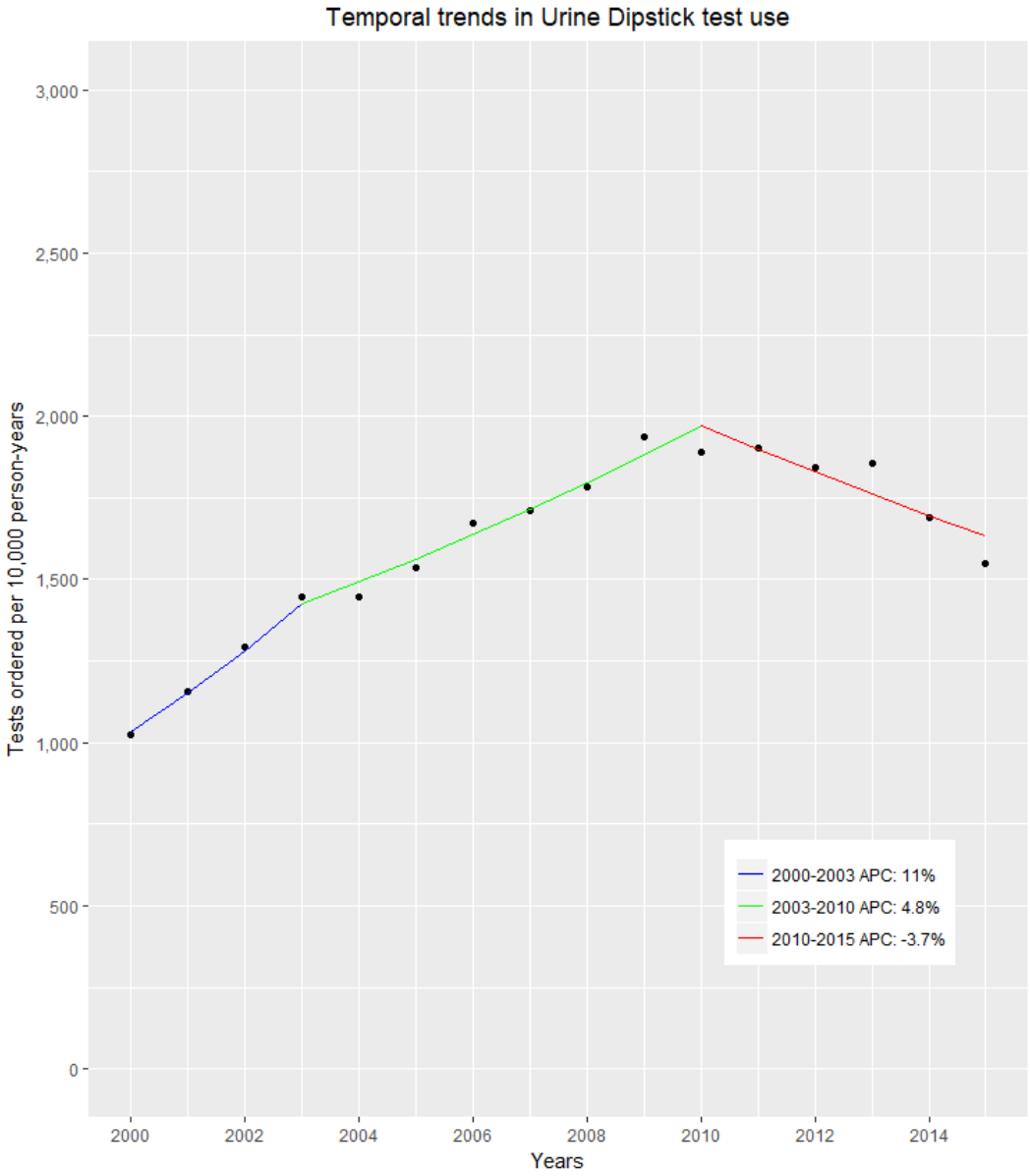


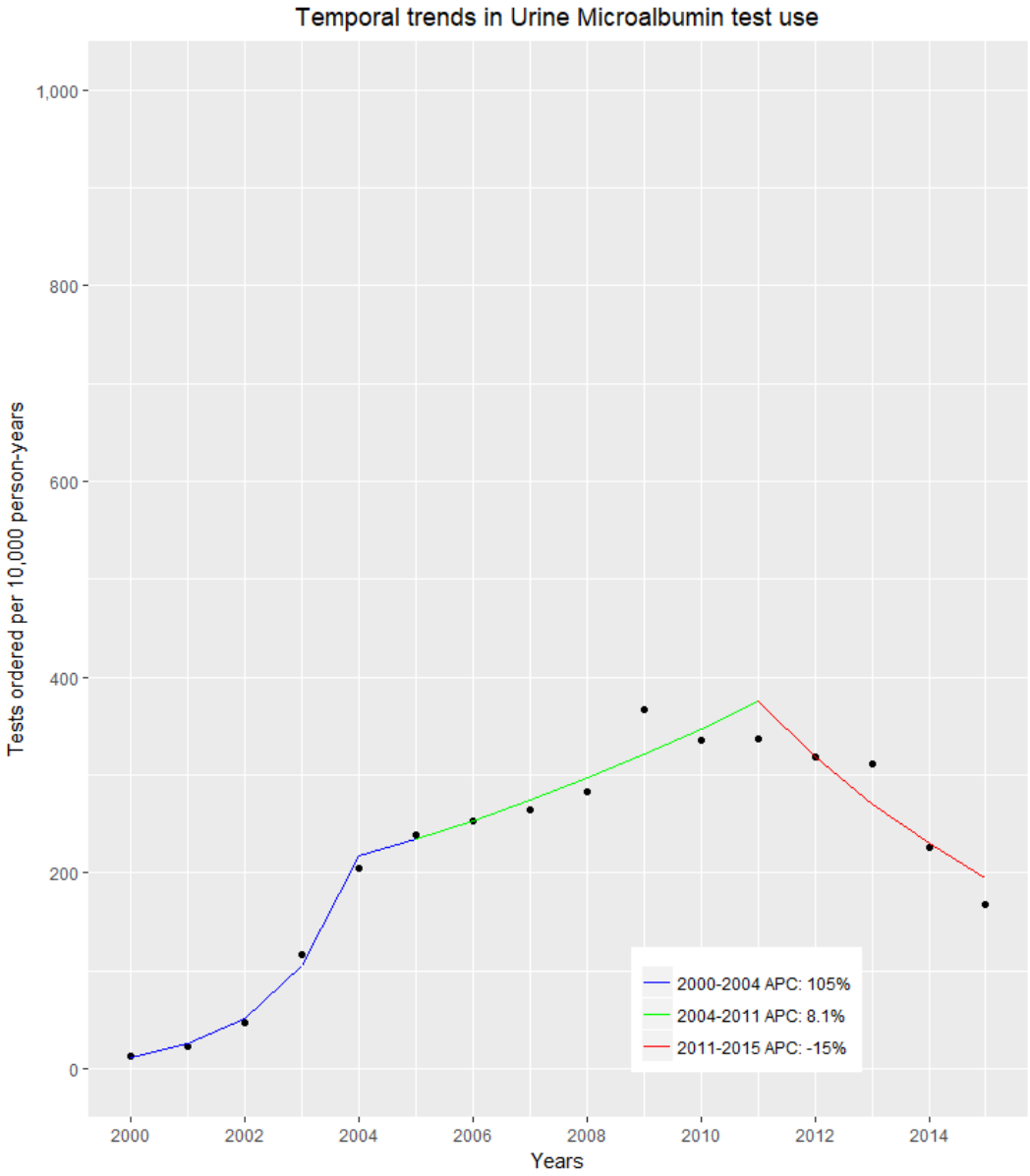
Pattern: Inverted U distribution

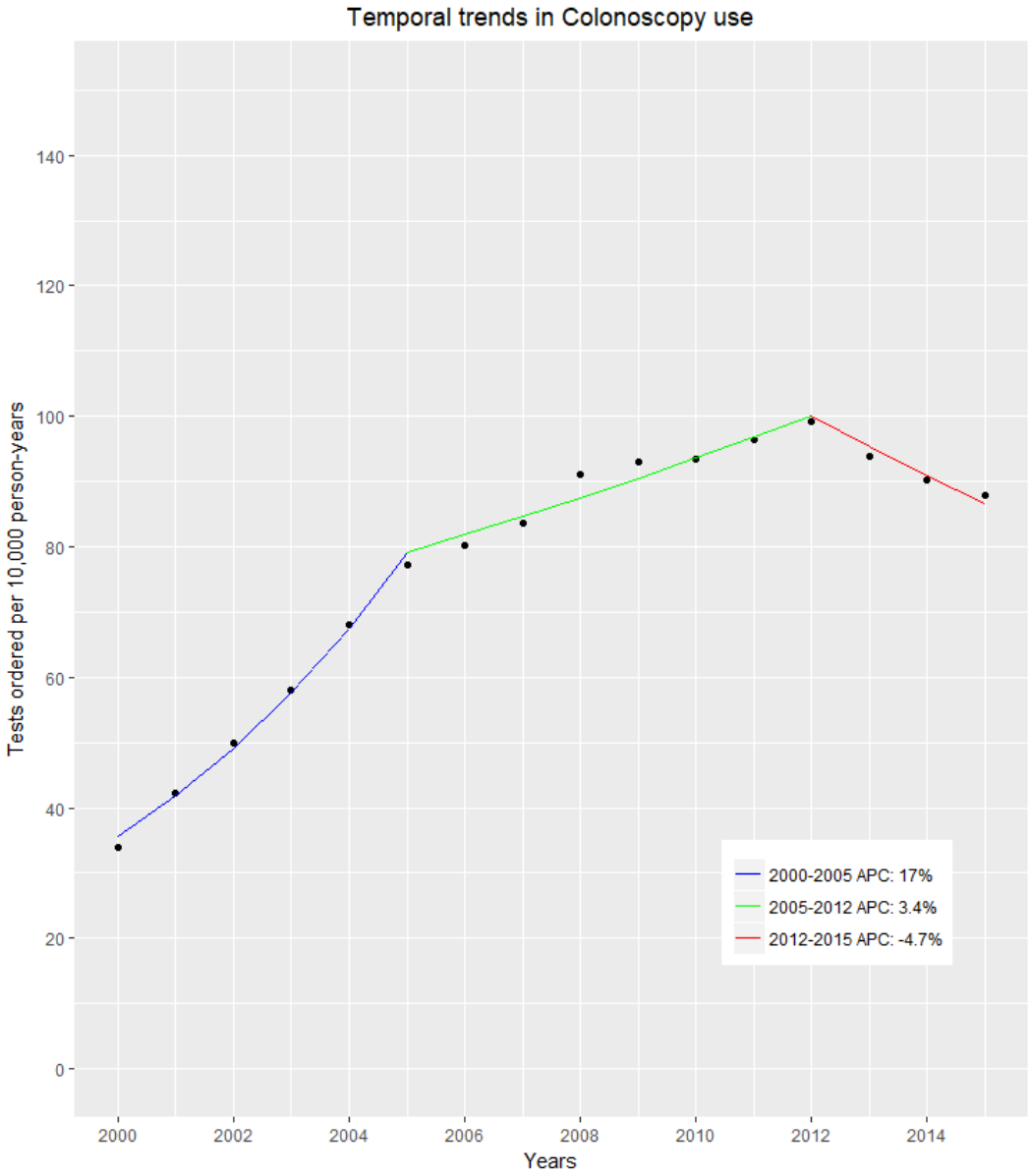

Temporal trends in Glucose test use
Tests ordered per 10,000 person-years
3,500
3,000
2,500
2,000
1,500
1,000
500
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 21%
2004-2011 APC: 3.0%
2011-2015 APC: -7.4%

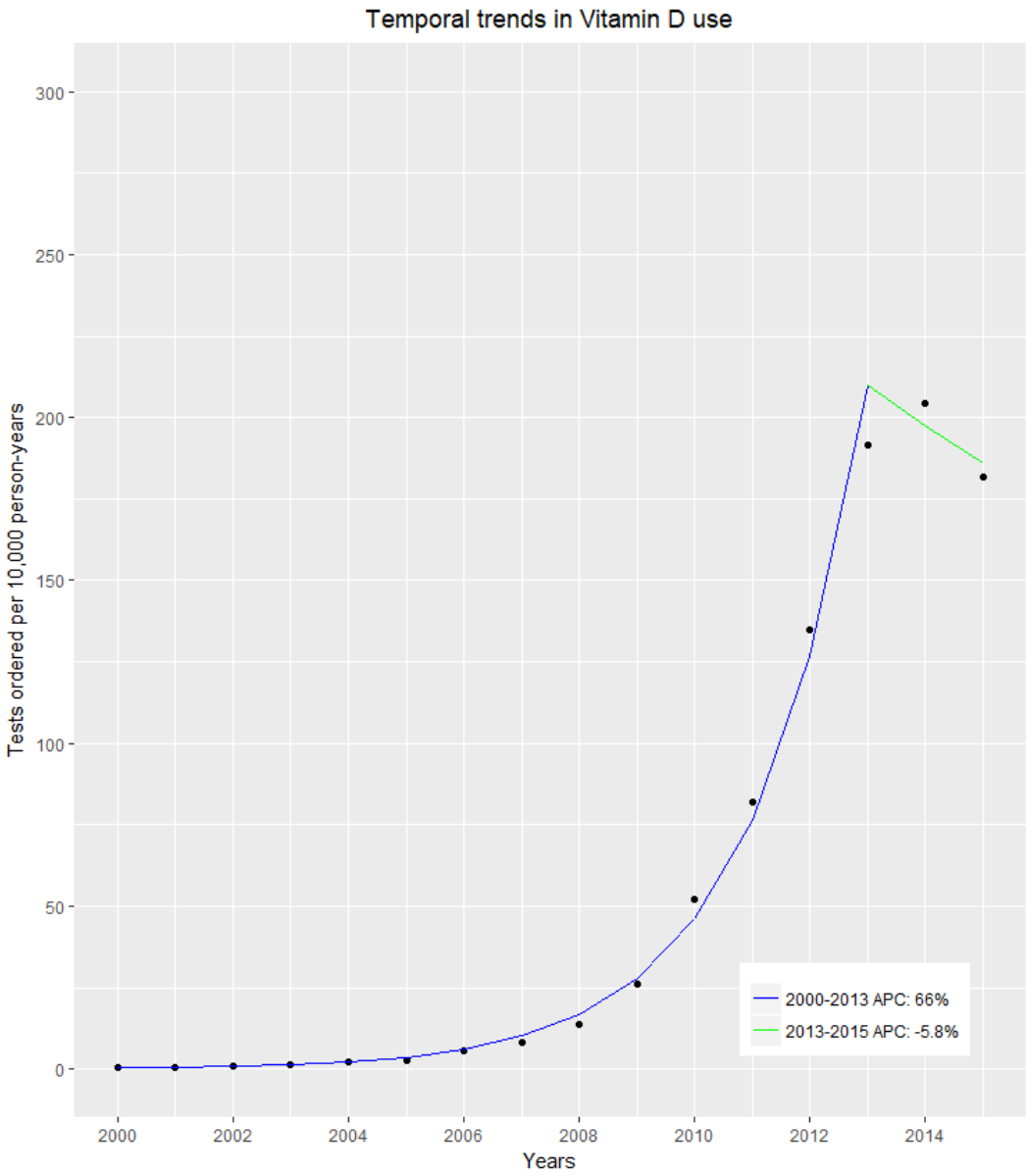

Temporal trends in Lipid test use
Tests ordered per 10,000 person-years
4,000
3,000
2,000
1,000
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 27%
2004-2013 APC: 1.4%
2013-2015 APC: -5.1%

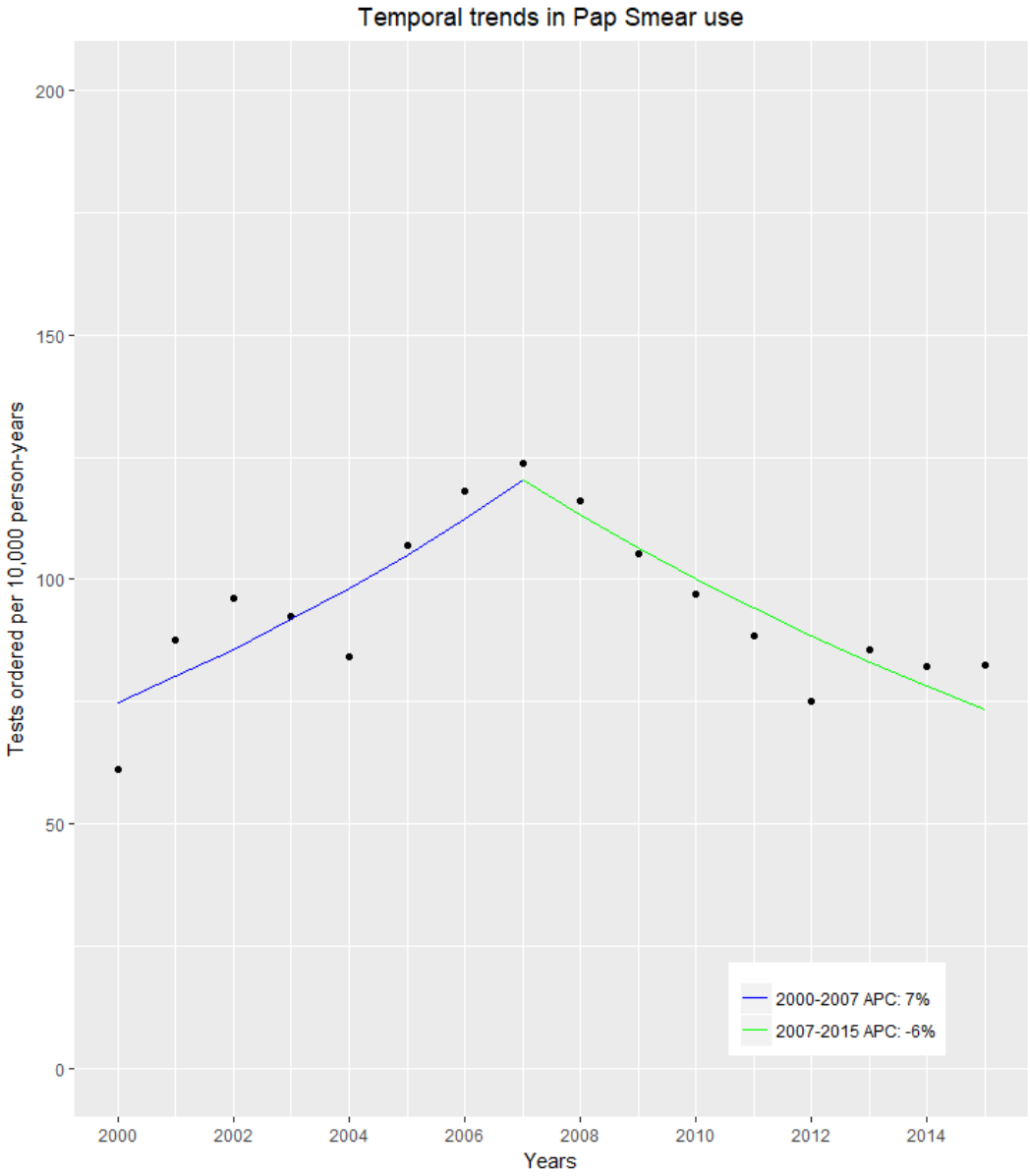

Temporal trends in Troponin test use
Tests ordered per 10,000 person-years
60
50
40
30
20
10
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 87%
2004-2011 APC: 4.0%
2011-2015 APC: -4.5%

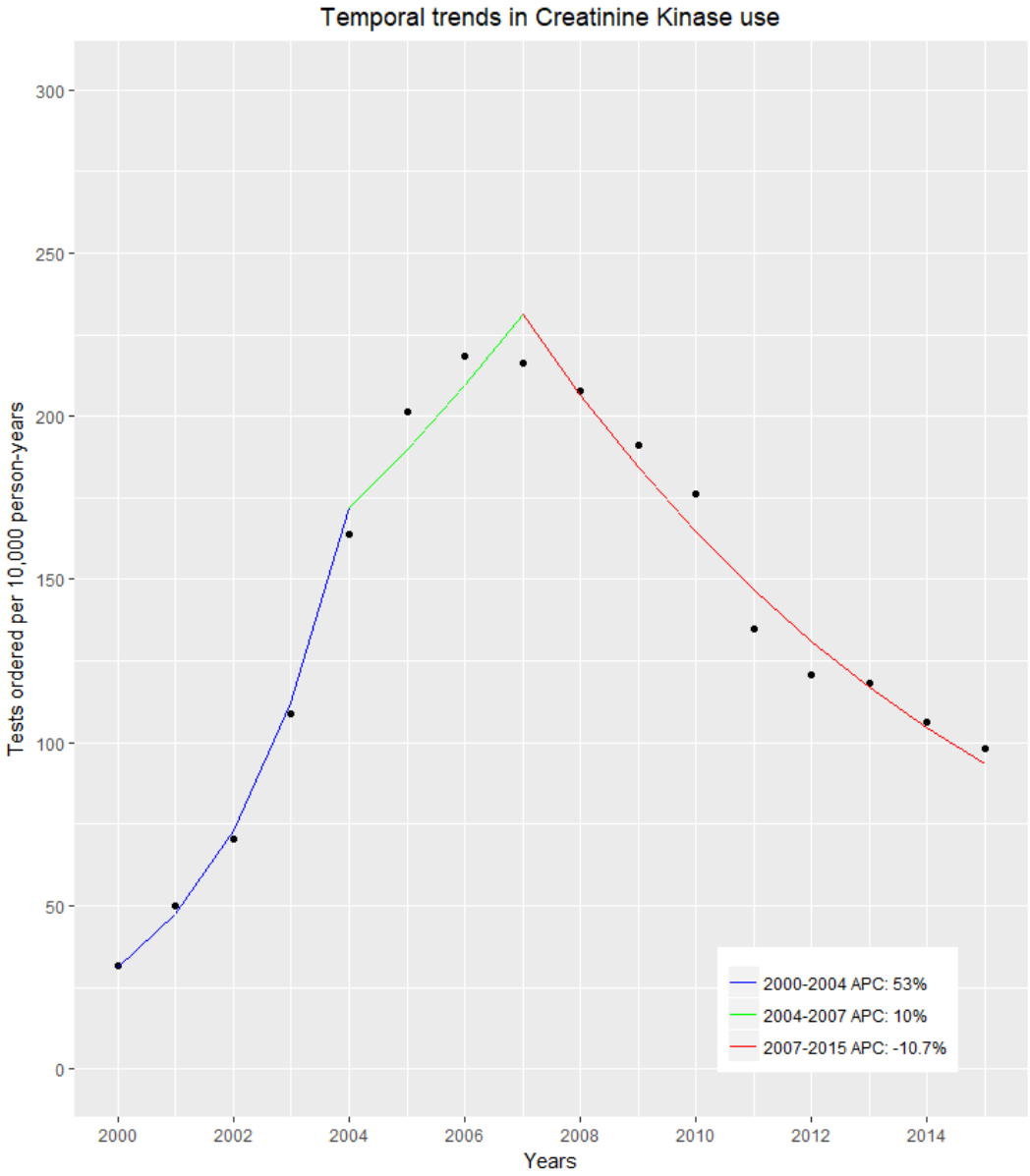

Temporal trends in Urine Dipstick test use
Tests ordered per 10,000 person-years
3,000
2,500
2,000
1,500
1,000
500
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2003 APC: 11%
2003-2010 APC: 4.8%
2010-2015 APC: -3.7%

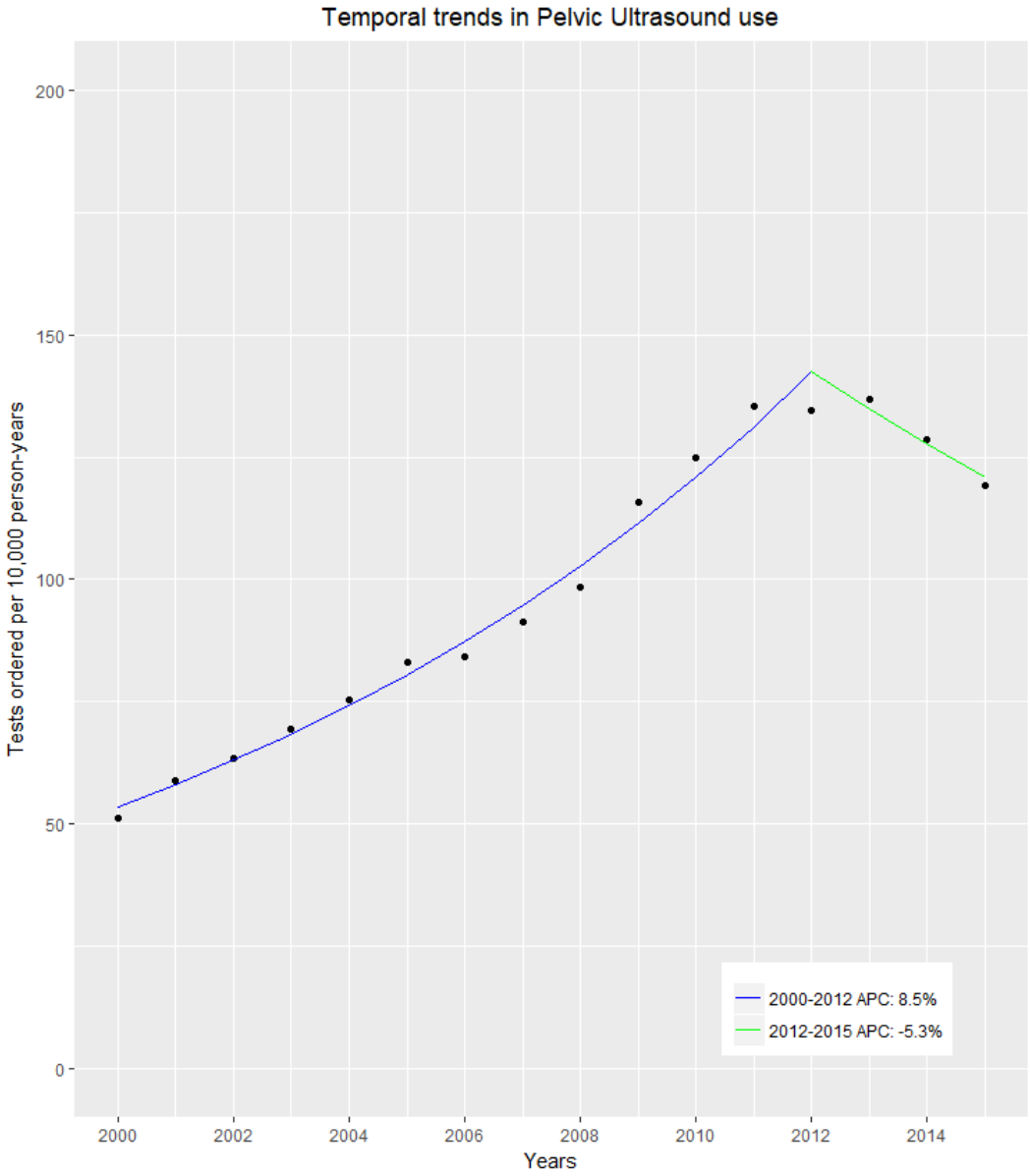

Temporal trends in Urine Microalbumin test use
Tests ordered per 10,000 person-years
1,000
800
600
400
200
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 105%
2004-2011 APC: 8.1%
2011-2015 APC: -15%


Temporal trends in Colonoscopy use
Tests ordered per 10,000 person-years
140
120
100
80
60
40
20
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2005 APC: 17%
2005-2012 APC: 3.4%
2012-2015 APC: -4.7%

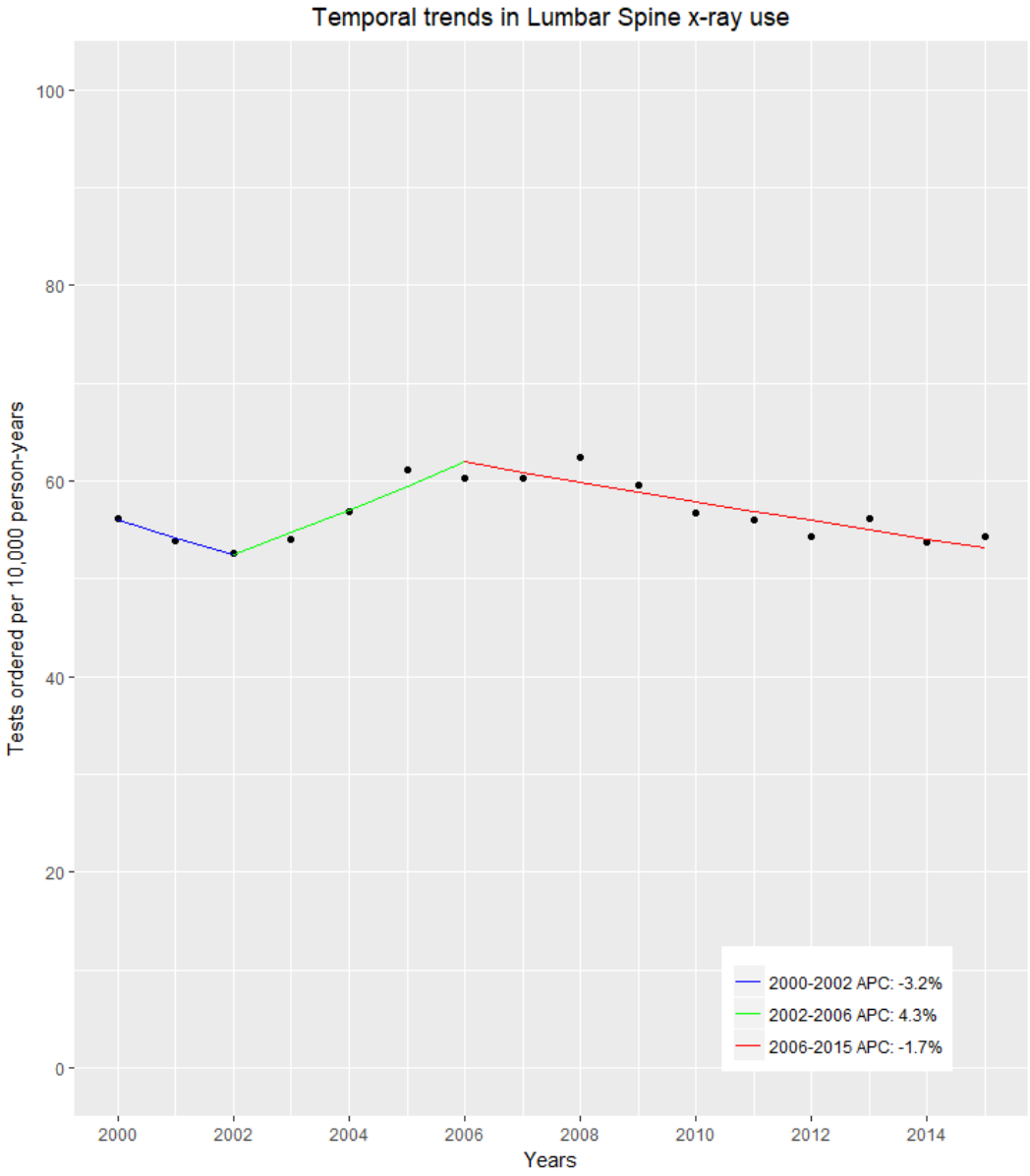

Temporal trends in Vitamin D use
Tests ordered per 10,000 person-years
300
250
200
150
100
50
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2013 APC: 66%
2013-2015 APC: -5.8%

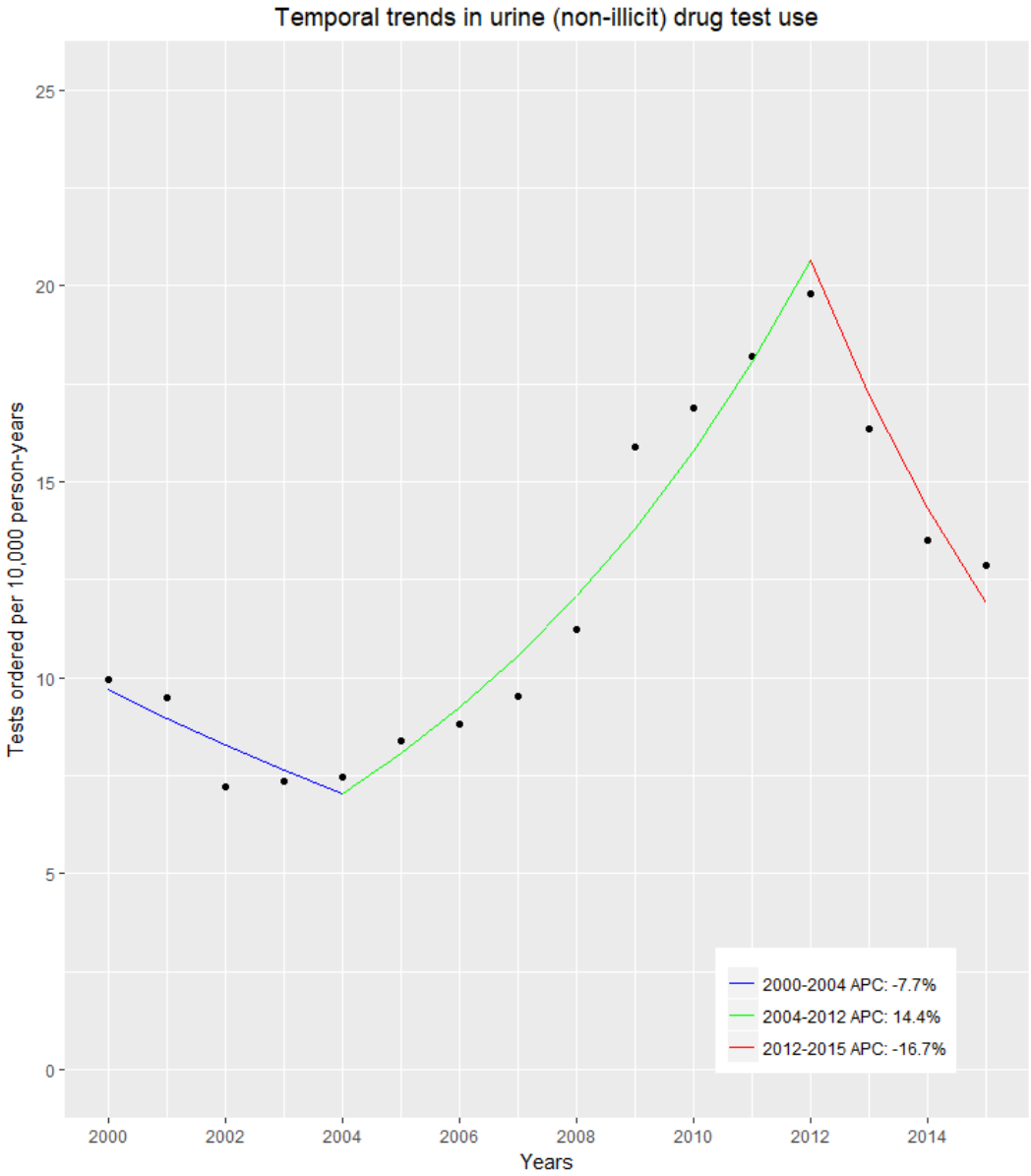

Temporal trends in Pap Smear use
Tests ordered per 10,000 person-years
200
150
100
50
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2007 APC: 7%
2007-2015 APC: -6%


Temporal trends in Creatinine Kinase use
Tests ordered per 10,000 person-years
300
250
200
150
100
50
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 53%
2004-2007 APC: 10%
2007-2015 APC: -10.7%

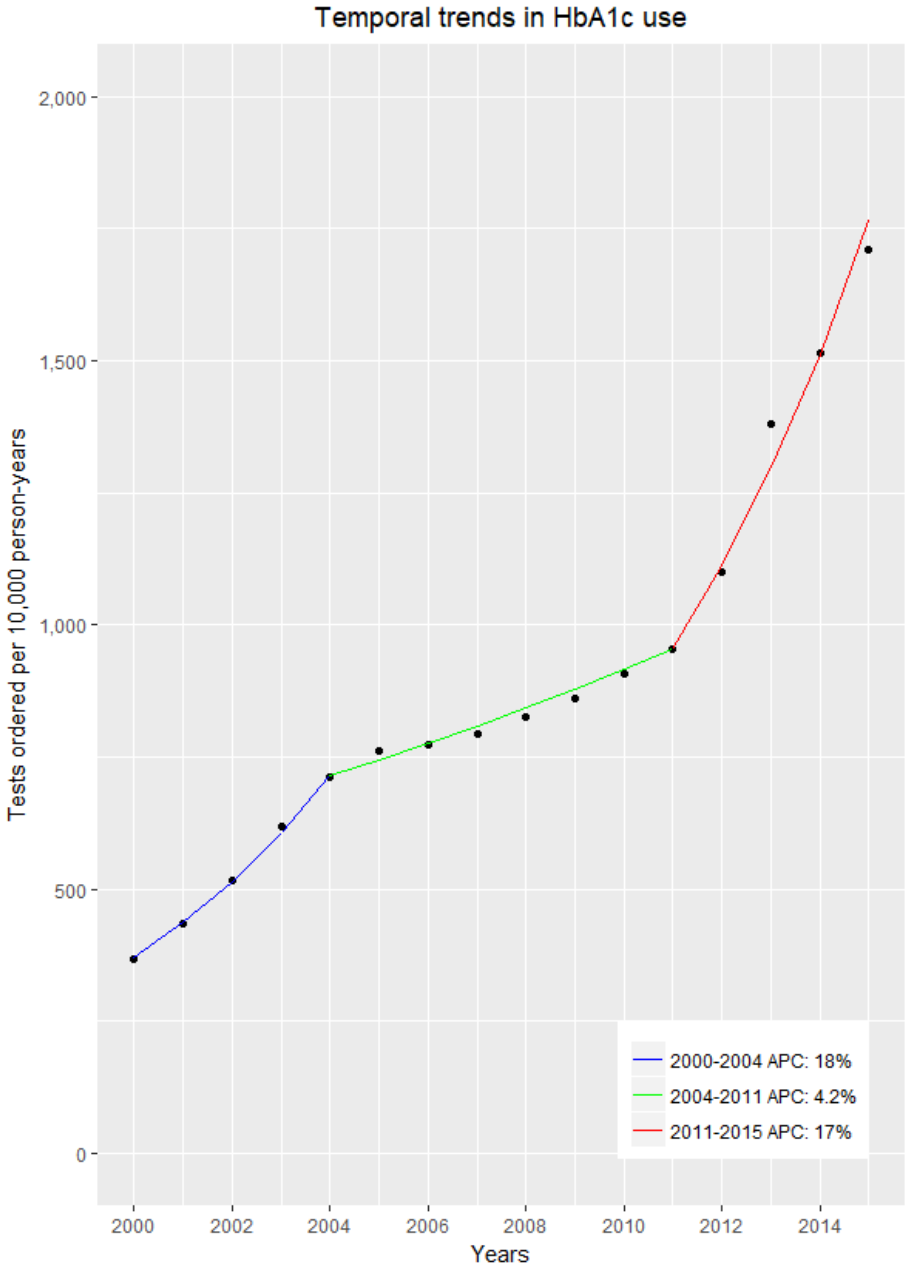


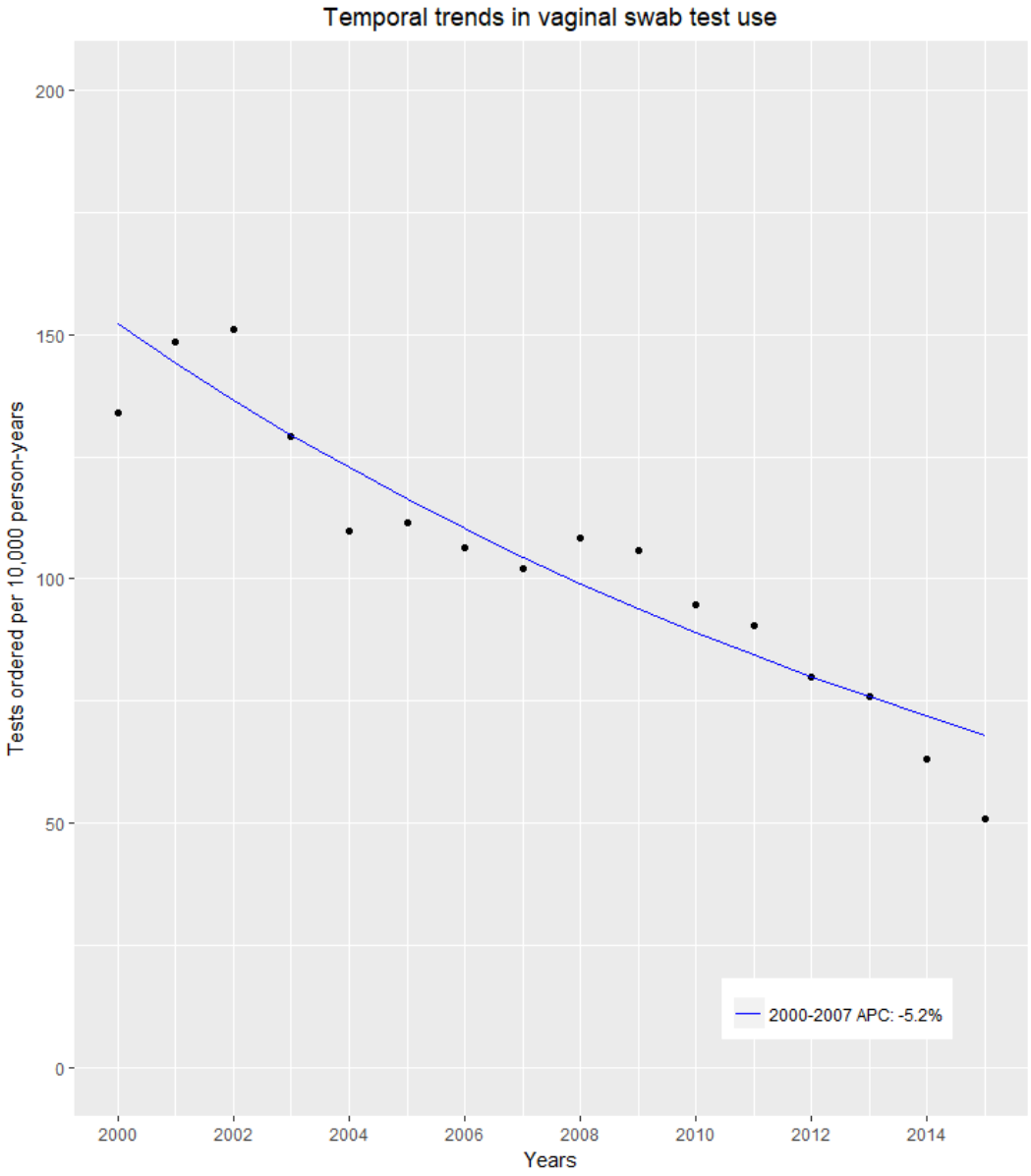
Pattern: A fall, an increase and then a fall


Temporal trends in Lumbar Spine x-ray use
Tests ordered per 10,000 person-years
100
80
60
40
20
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2002 APC: -3.2%
2002-2006 APC: 4.3%
2006-2015 APC: -1.7%



Pattern: Exceptions


Temporal trends in HbA1c use
Tests ordered per 10,000 person-years
2,000
1,500
1,000
500
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2004 APC: 18%
2004-2011 APC: 4.2%
2011-2015 APC: 17%


Temporal trends in vaginal swab test use
Tests ordered per 10,000 person-years
200
150
100
50
0
2000
2002
2004
2006
2008
2010
2012
2014
Years
2000-2007 APC: -5.2%

3.3 RECORD Checklist

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items	Location in manuscript where items are reported
Title and abstract					
	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found		RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included. RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract. RECORD 1.3: If linkage between	Title and abstract: page 37

				databases was conducted for the study, this should be clearly stated in the title or abstract.	
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported			Introduction: Page 37
Objectives	3	State specific objectives, including any prespecified hypotheses			Introduction: Page 37
Methods					
Study Design	4	Present key elements of study design early in the paper			Methods: Page 38
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection			Methods: Page 38
Participants	6	*(a) Cohort study* - Give the		RECORD 6.1: The methods of study	Methods: Page 38-

		eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study* - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study* - Give the eligibility criteria, and the sources and methods of selection of participants *(b) Cohort study* - For matched studies, give matching criteria and		population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided. RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process,	39

		number of exposed and unexposed *Case-control study* - For matched studies, give matching criteria and the number of controls per case		including the number of individuals with linked data at each stage.	
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.		RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	Methods: Page 38-39
Data sources/ measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group			Methods: Page 38-39

Bias	9	Describe any efforts to address potential sources of bias			Methods: Page 38-39
Study size	10	Explain how the study size was arrived at			Methods: Page 38-39
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why			Methods: Page 38-39
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) *Cohort study* - If applicable,			Methods: Page 38-39

		explain how loss to follow-up was addressed *Case-control study* - If applicable, explain how matching of cases and controls was addressed *Cross-sectional study* - If applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses			
Data access and cleaning methods		..		RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population. RECORD 12.2: Authors should provide	Methods: Page 38-39

				information on the data cleaning methods used in the study.	
Linkage		..		RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.	Methods: Page 38-39
Results					
Participants	13	(a) Report the numbers of individuals at each stage of the study (*e.g.*, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed) (b) Give reasons for non-		RECORD 13.1: Describe in detail the selection of the persons included in the study (*i.e.,* study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.	Results: Page 40 - 48

		participation at each stage. (c) Consider use of a flow diagram			
Descriptive data	14	(a) Give characteristics of study participants (*e.g.*, demographic, clinical, social) and information on exposures and potential confounders (b) Indicate the number of participants with missing data for each variable of interest (c) *Cohort study* - summarise follow-up time (*e.g.*, average and total amount)			Results: Page 40 - 48
Outcome data	15	*Cohort study* - Report numbers of outcome events or summary measures over time *Case-control study* - Report			Results: Page 40 - 48

		numbers in each exposure category, or summary measures of exposure *Cross-sectional study* - Report numbers of outcome events or summary measures			
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorized (c) If relevant, consider translating			Results: Page 40 - 48

		estimates of relative risk into absolute risk for a meaningful time period			
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses			Results: Page 40 – 48 and appendix
Discussion					
Key results	18	Summarise key results with reference to study objectives			Discussion page 49 - 54
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias		RECORD 19.1: Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to	Discussion page 49 - 54

				the study being reported.	
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence			Discussion page 49 - 54
Generalisability	21	Discuss the generalisability (external validity) of the study results			Discussion page 49 - 54
Other Information					
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based			Page 38
Accessibility of		..		RECORD 22.1: Authors should provide	Methods page 3

protocol, raw data, and programming code				information on how to access any supplemental information such as the study protocol, raw data, or programming code.	

Appendix Chapter 4

4.1 Included tests

Below we list the 44 tests grouped into test type (28 laboratory, 11 imaging and five other, miscellaneous tests)

Laboratory tests

- Non-Illicit Drug monitoring
- Urine Albumin
- Troponin
- Clotting
- Vitamin D
- Creatinine Kinases
- Prostate specific Antigen (PSA)
- Iron
- Progesterone
- Bone Profile
- Creatine Reactive Protein (CRP)
- Urine Microscopy culture sensitivities (MCS)
- Oestradiol
- Urine Albumin: Creatinine Ratio (ACR)
- HbA1c
- Vitamin B12
- Erythrocyte Sedimentation Rate (ESR)
- Folate
- Liver Function Tests (LFT)
- Full Blood Count (FBC)

Imaging tests

- Pelvic computer tomography (CT)
- Brain Magnetic resonance imaging (MRI)
- Knee MRI
- Dual-energy X-ray absorptiometry (DEXA)
- Lumbar Spine MRI
- Brain CT
- Lumbar spine x-ray
- Chest x-rays
- Knee x-ray
- Pelvic Ultrasound

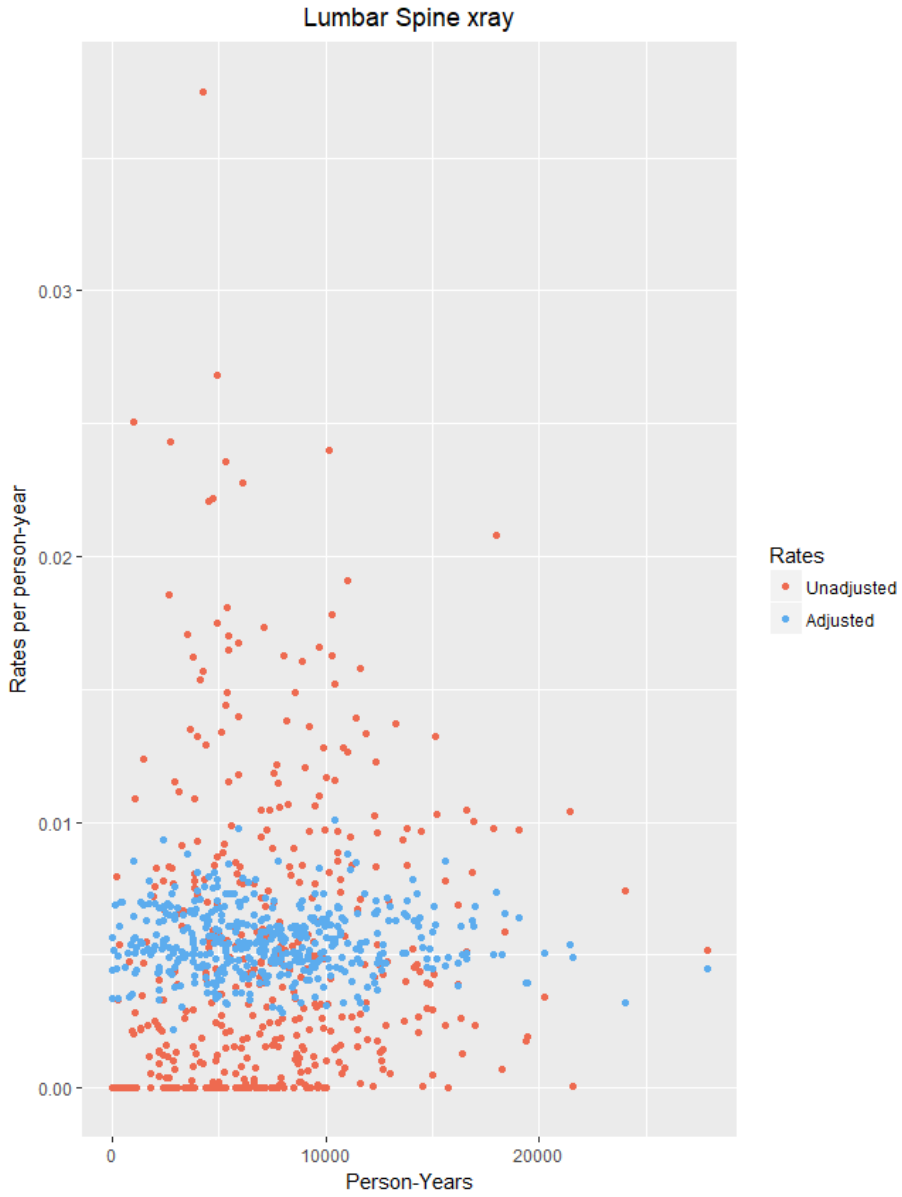
- Echocardiogram

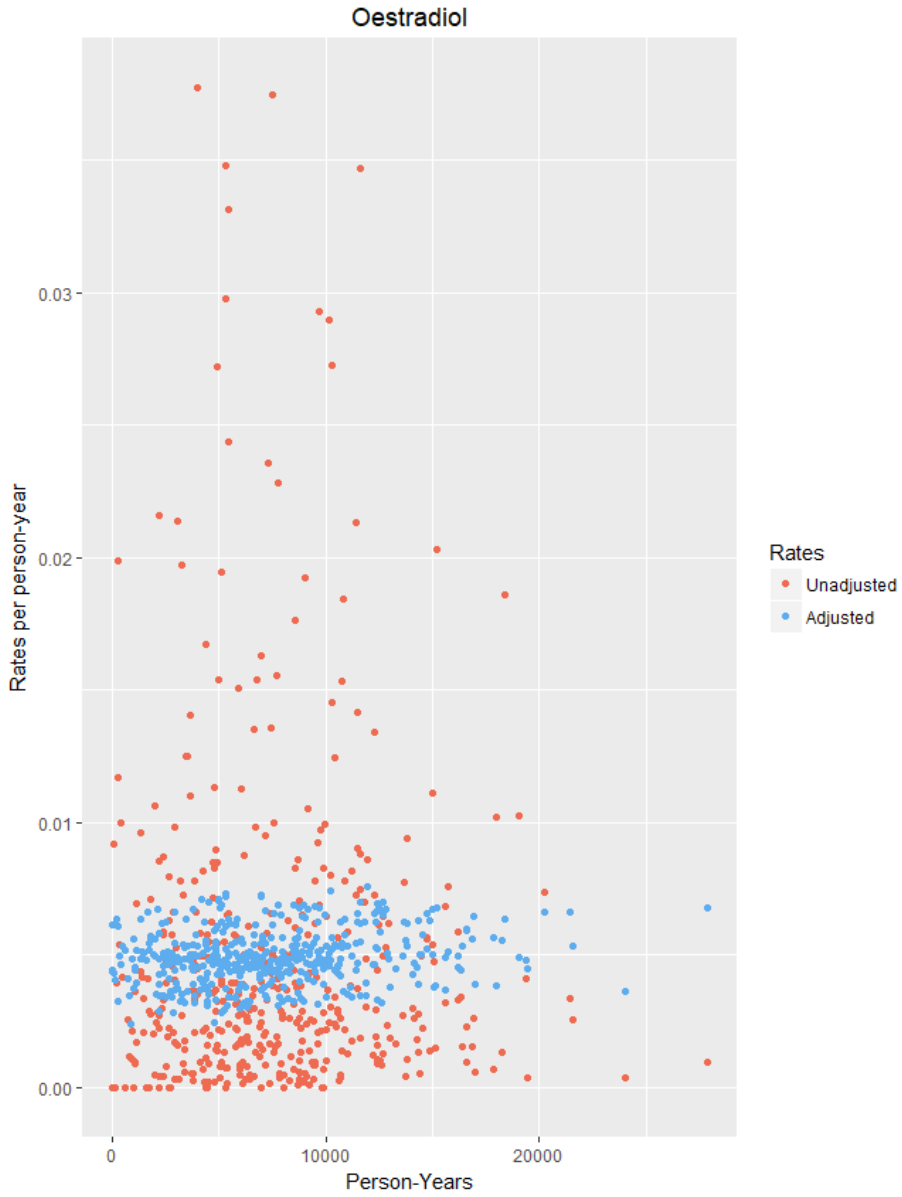
Other, miscellaneous tests

- Pap Smear
- Vaginal Swab
- Upper Endoscopy
- Colonoscopy

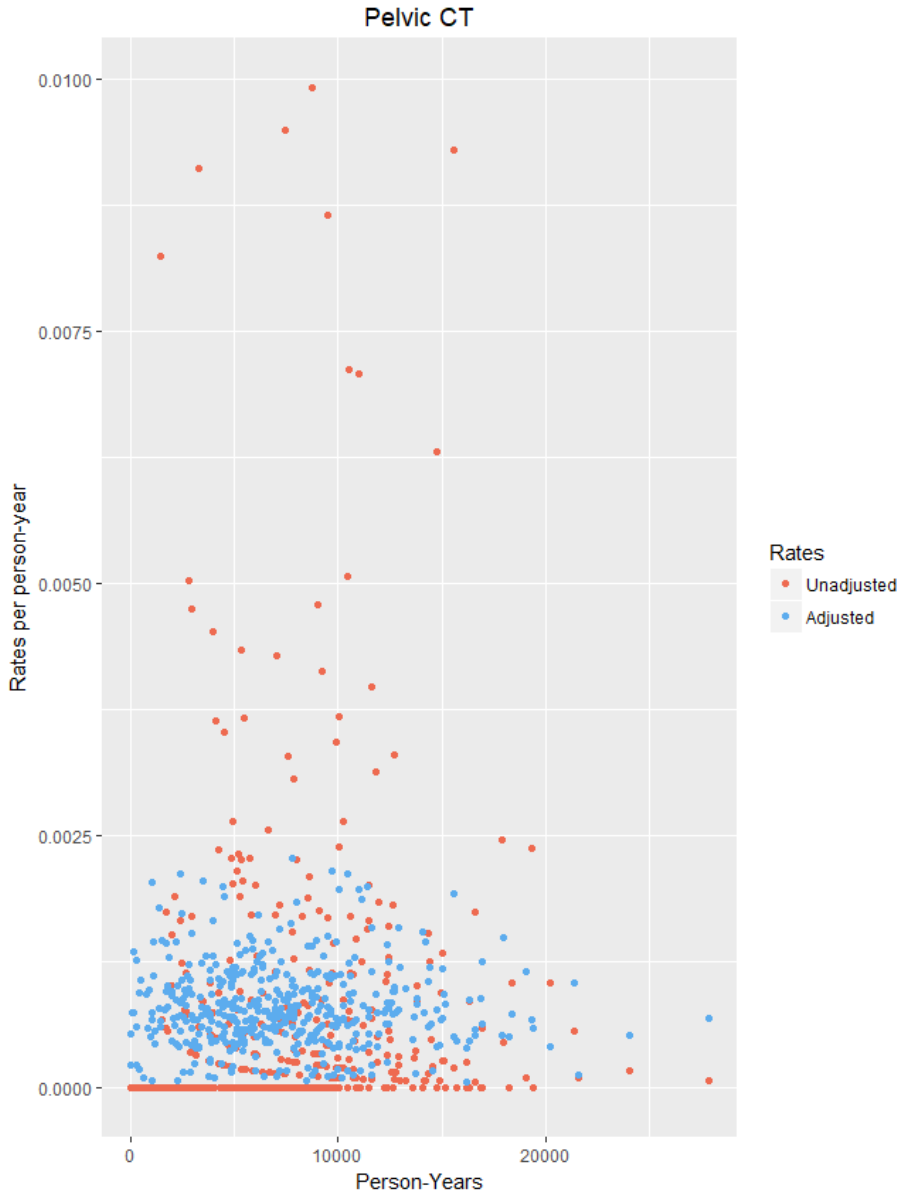
4.2 Adjusted and unadjusted coefficient of variation for each 44 tests (individual graphs)

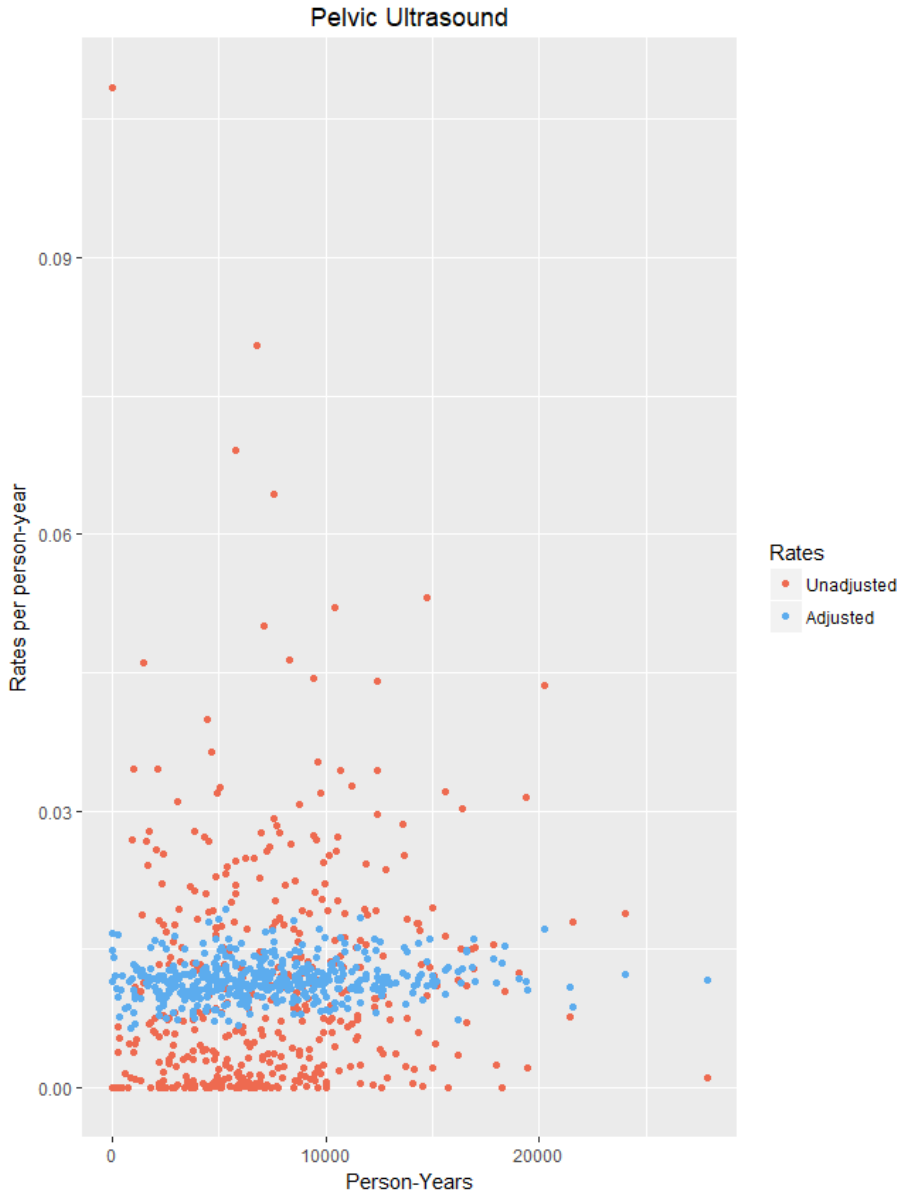

Lipids
Rates per person-year
0.6
0.4
0.2
0.0
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

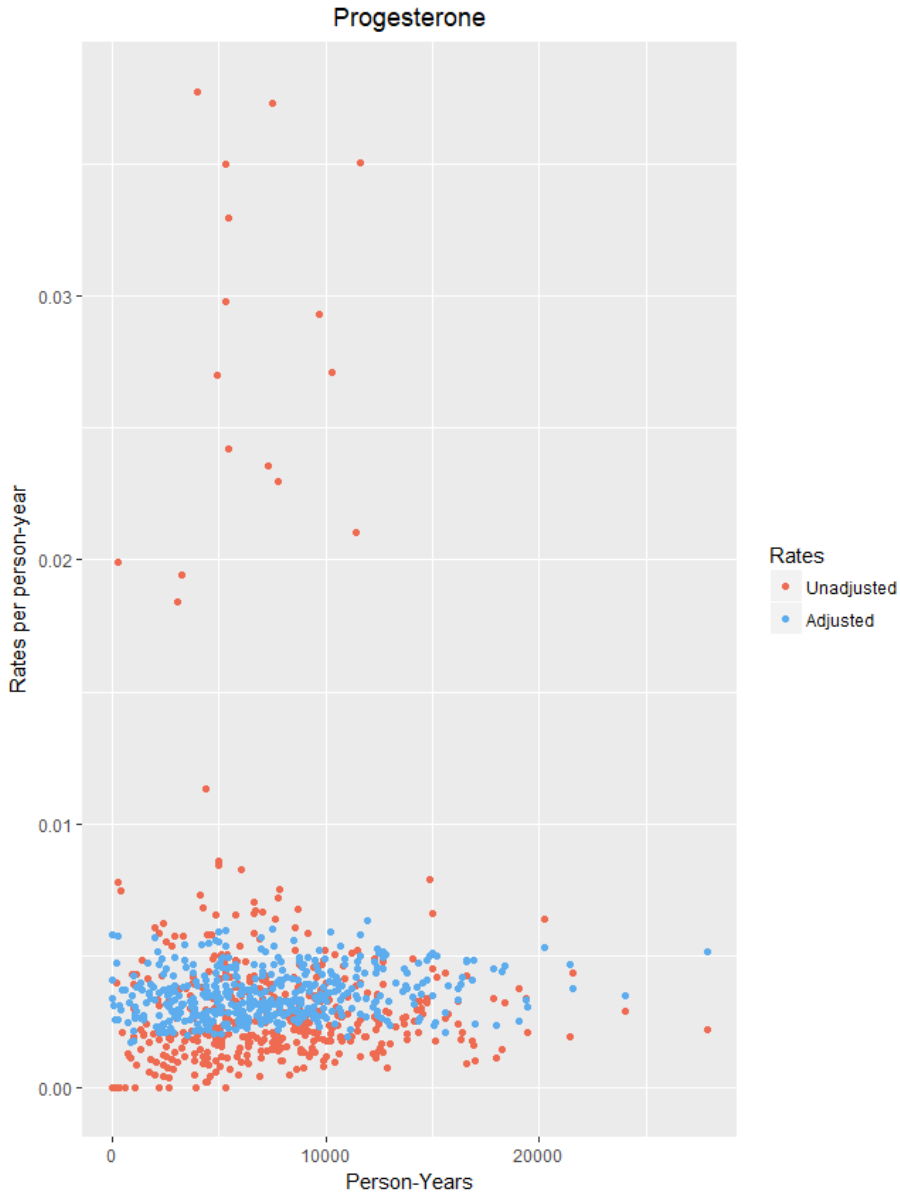

Lumbar Spine xray
Rates per person-year
0.03
0.02
0.01
0.00
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

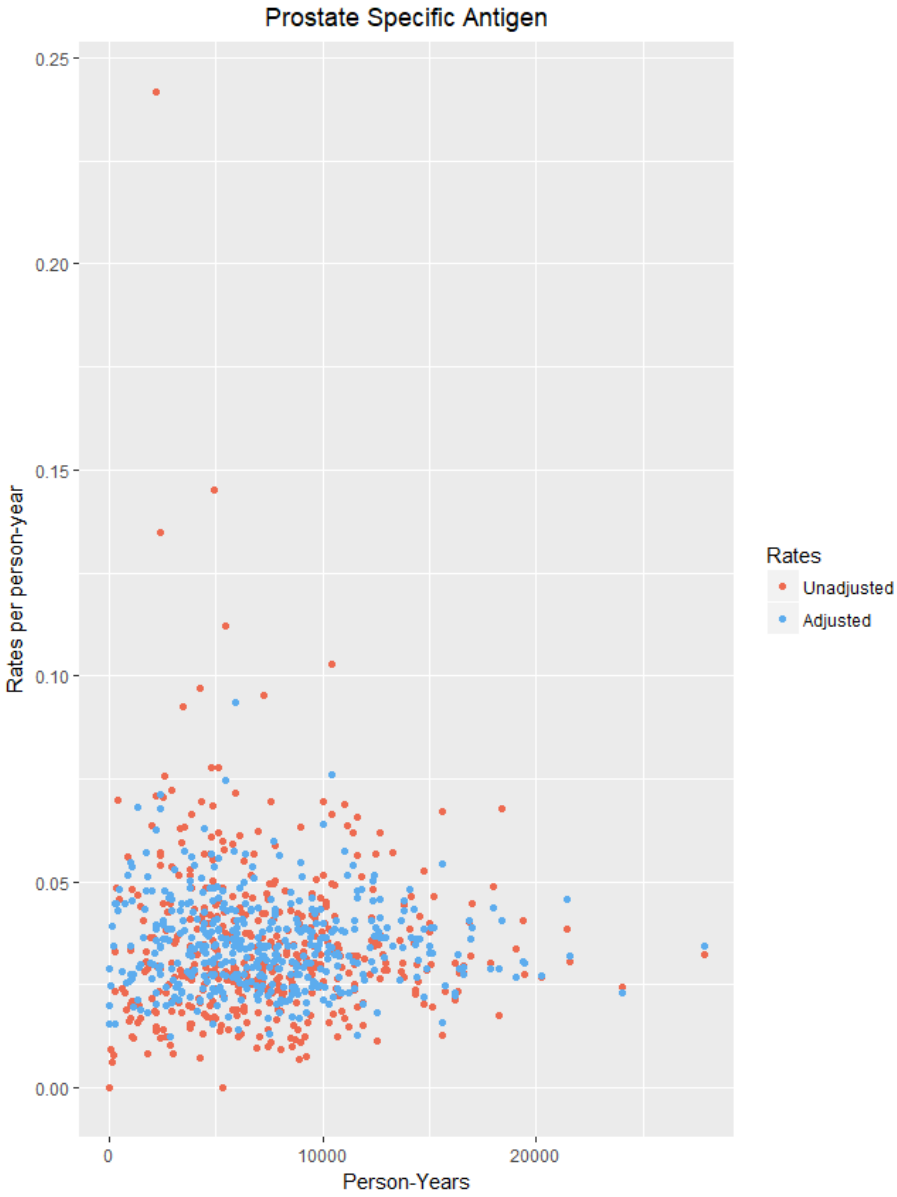

Oestradiol
Rates per person-year
0.03
0.02
0.01
0.00
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

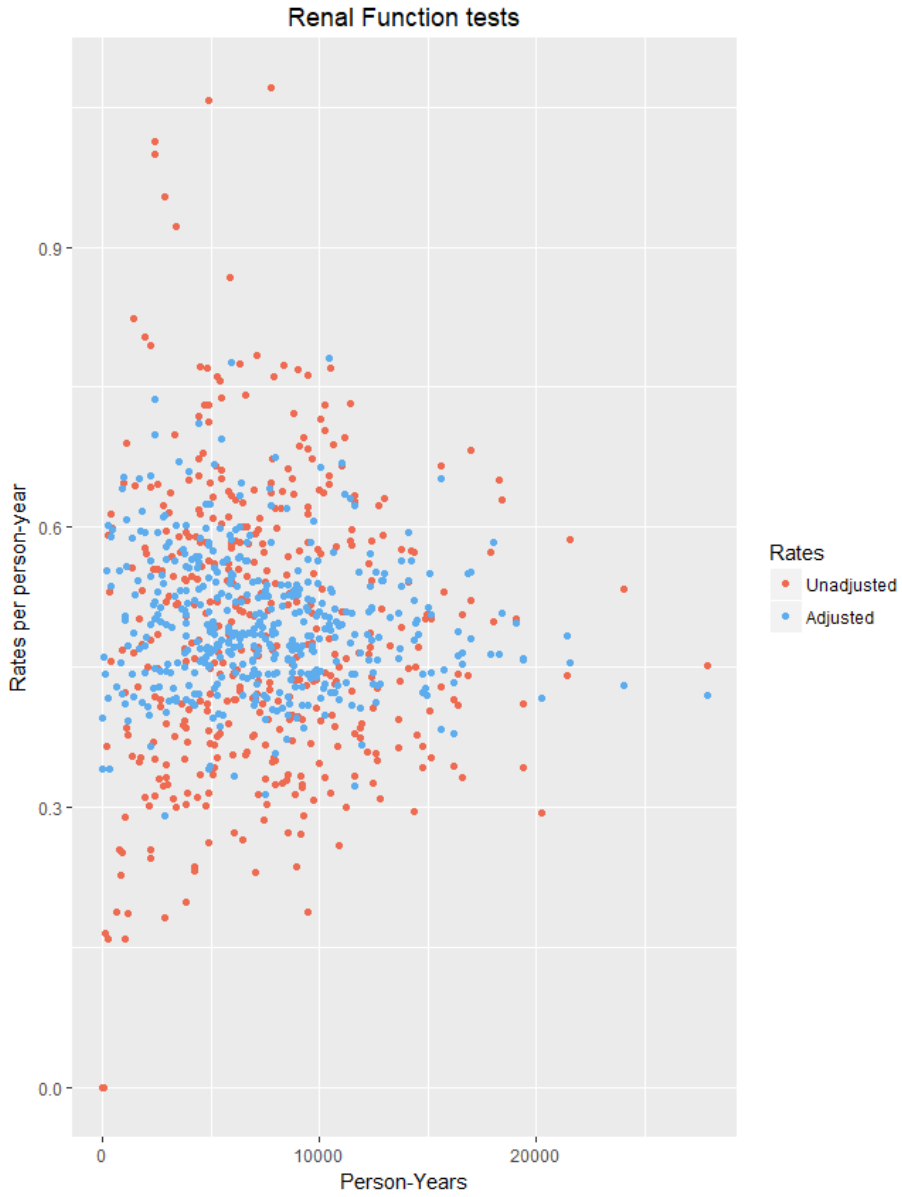

Pap Smear
Rates per person-year
0.06
0.04
0.02
0.00
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

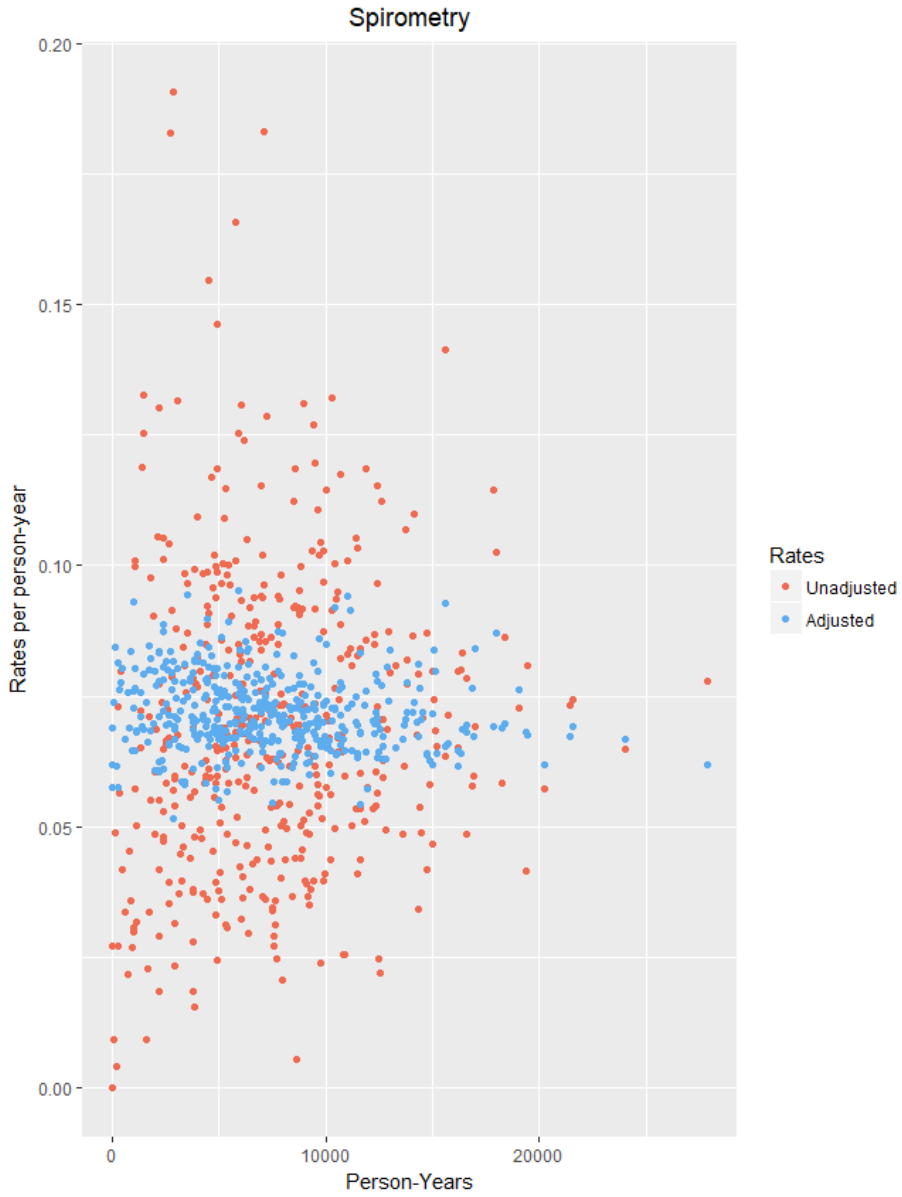

Pelvic CT
0.0100
0.0075
0.0050
0.0025
0.0000
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

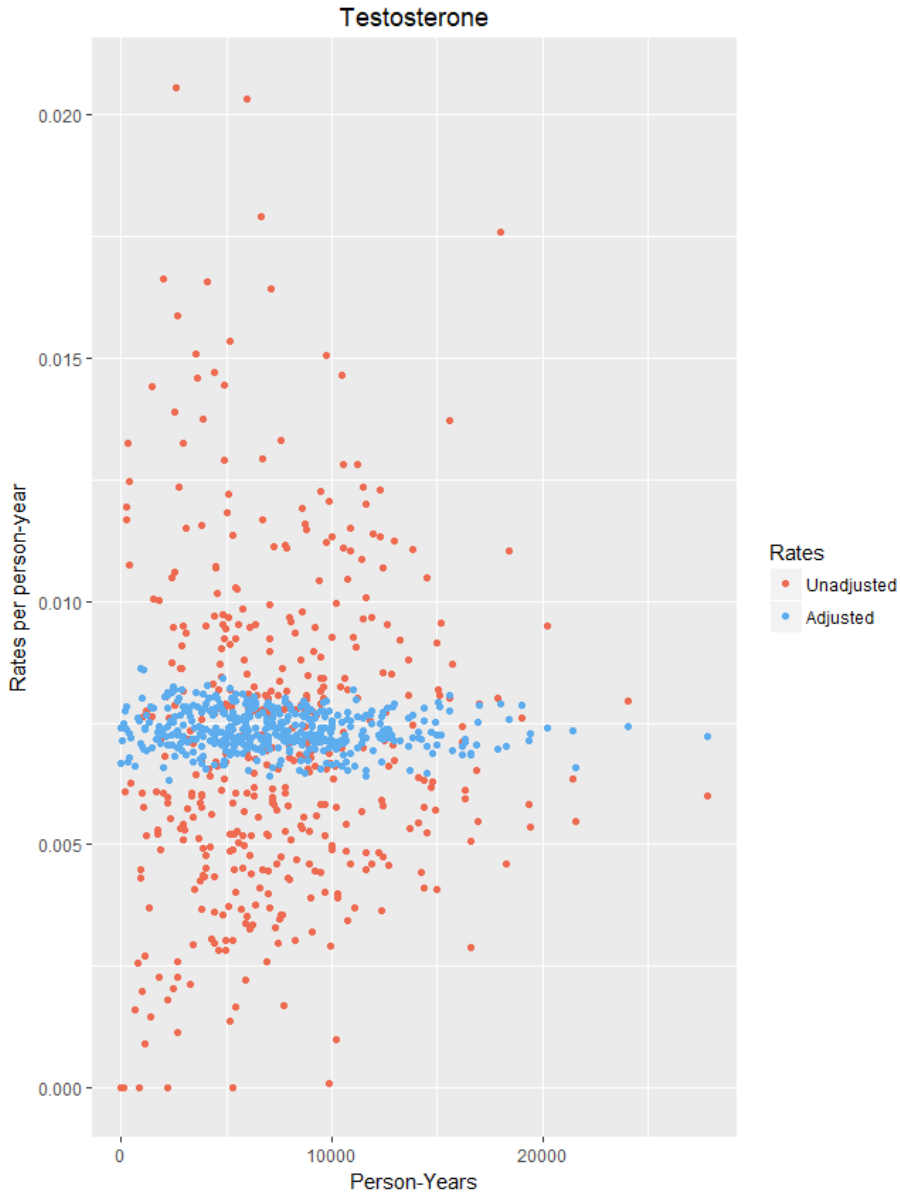

Pelvic Ultrasound
Rates per person-year
0.09
0.06
0.03
0.00
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

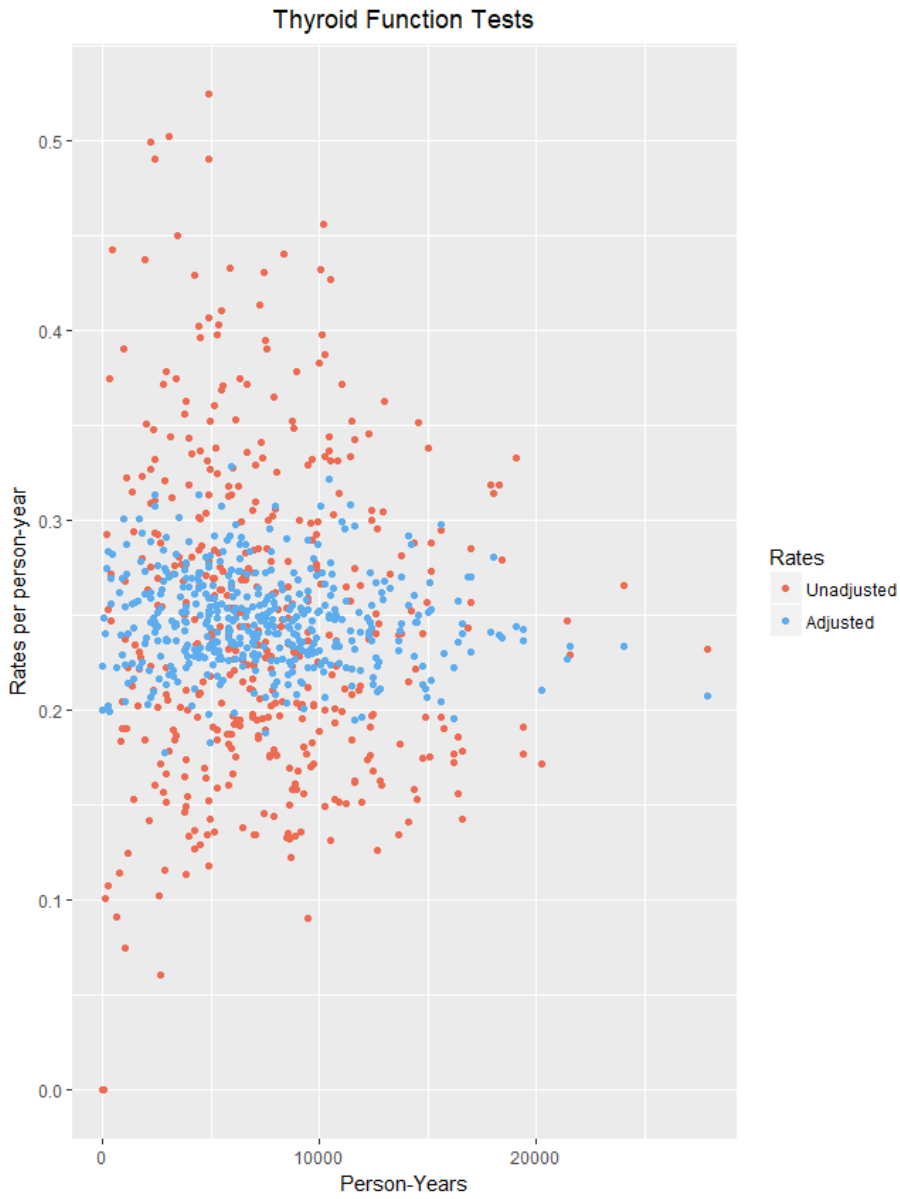

Progesterone
Rates per person-year
0.03
0.02
0.01
0.00
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

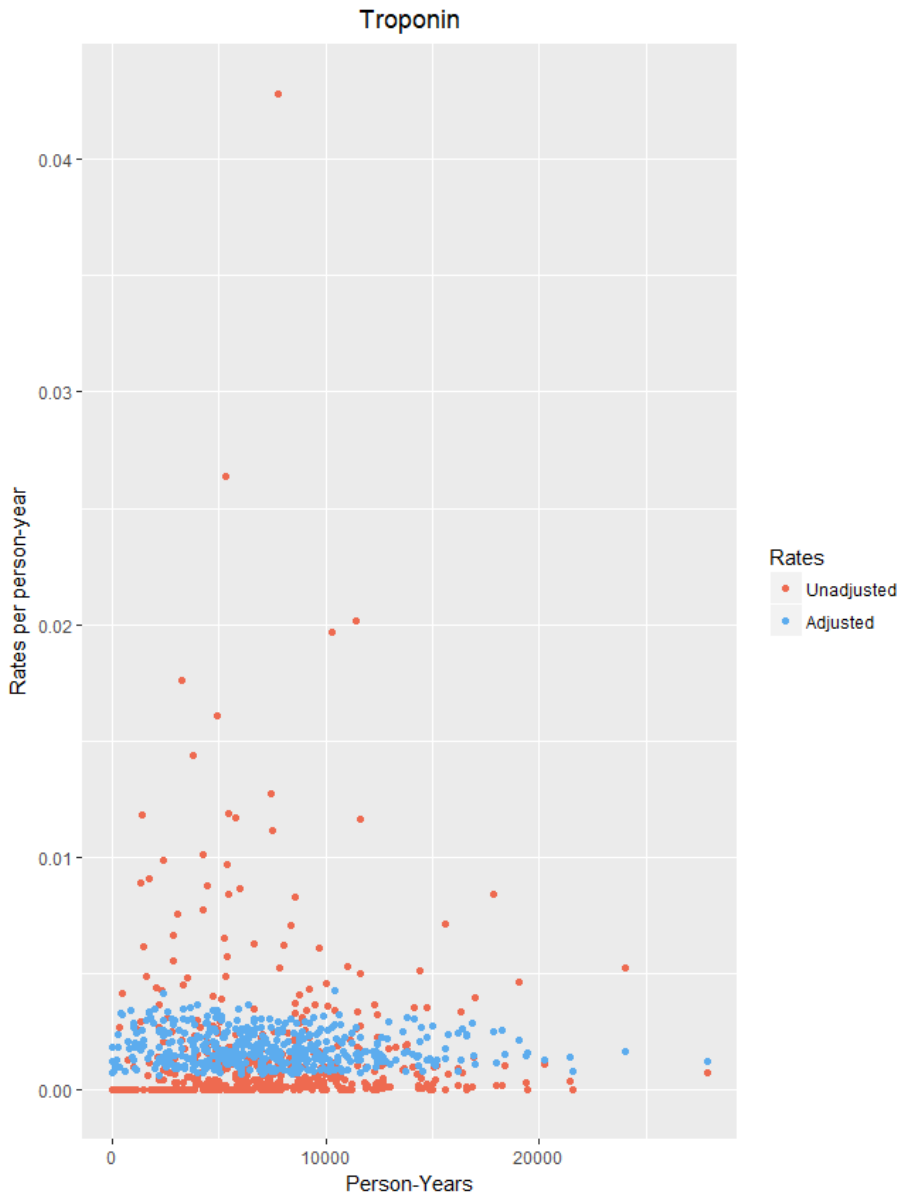

Prostate Specific Antigen
0.25
0.20
0.15
0.10
0.05
0.00
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted


Renal Function tests
Rates per person-year
0.0
0.3
0.6
0.9
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

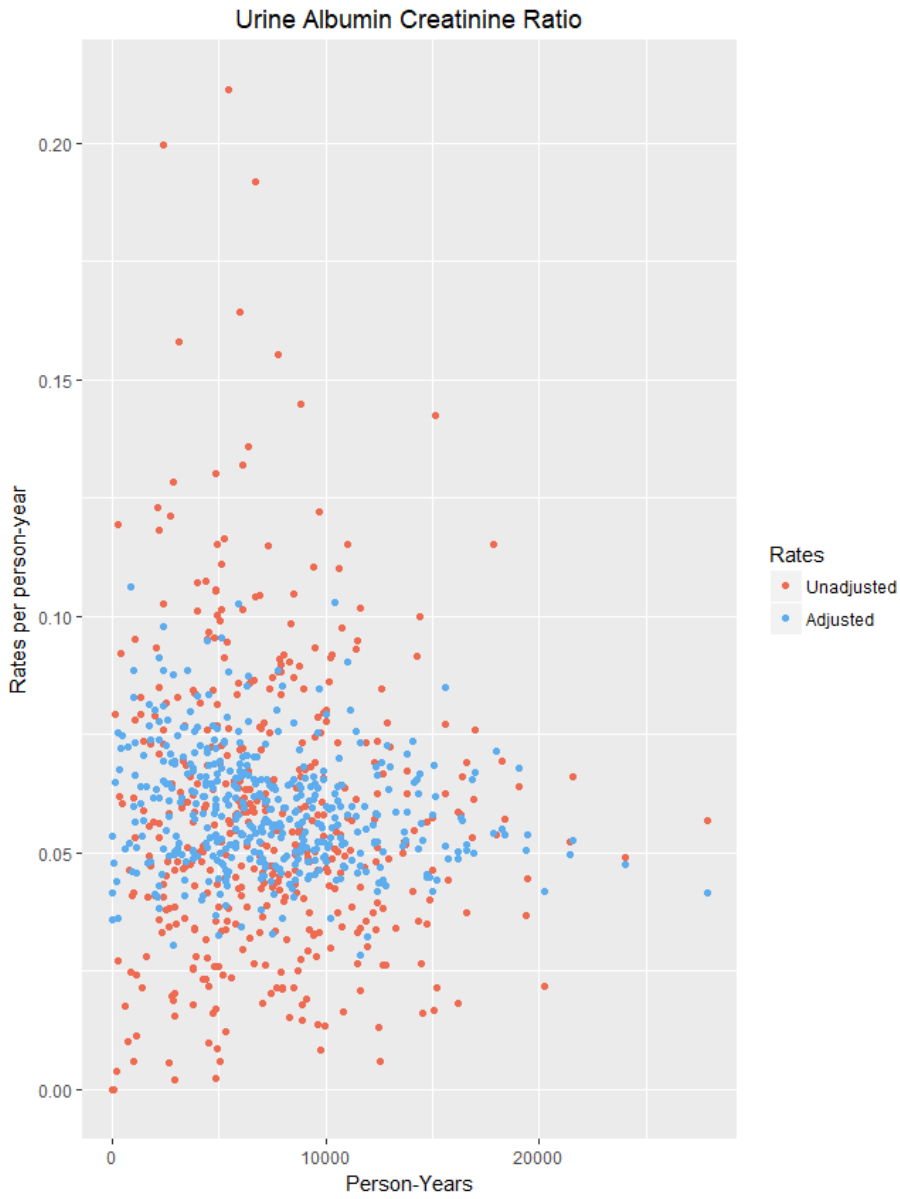

Spirometry
0.20
0.15
0.10
0.05
0.00
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

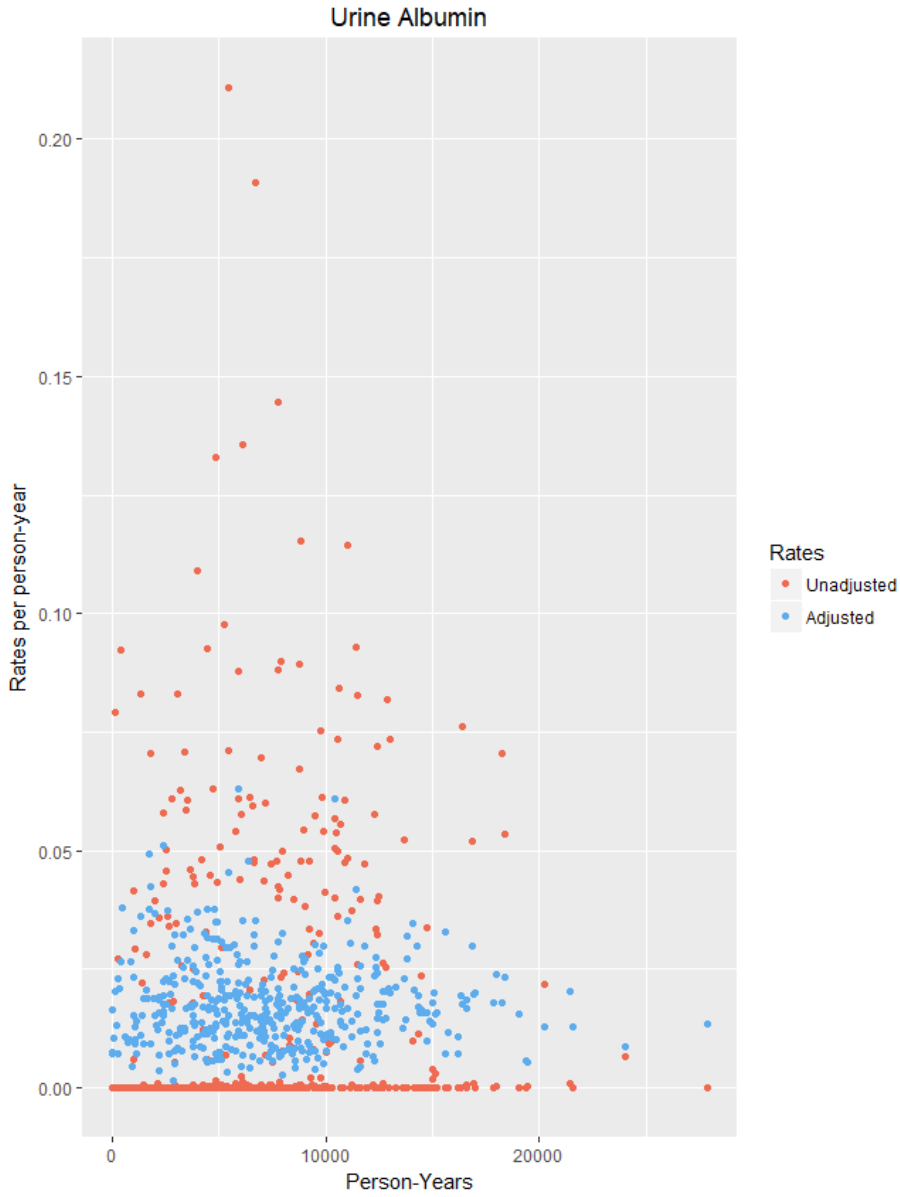

Testosterone
0.020
0.015
0.010
0.005
0.000
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

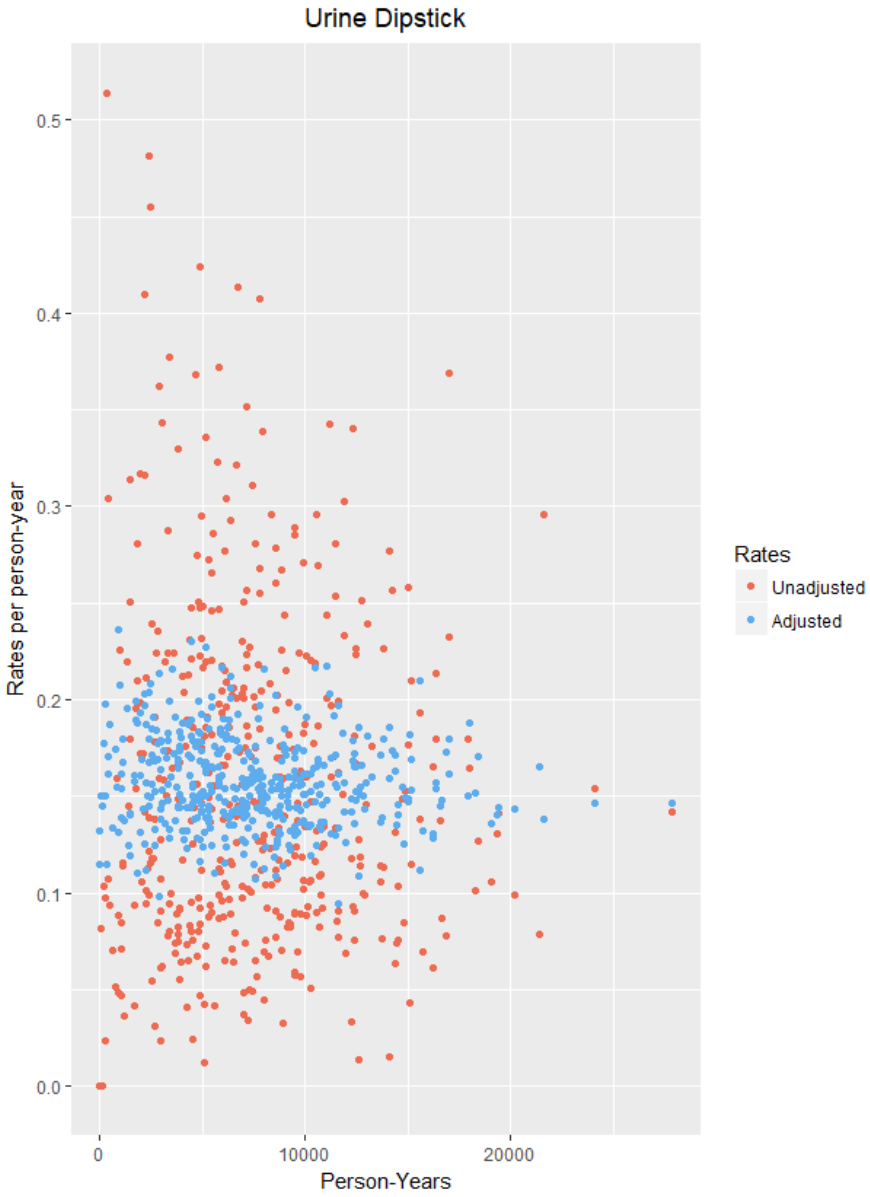

Thyroid Function Tests
Rates per person-year
0.5
0.4
0.3
0.2
0.1
0.0
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

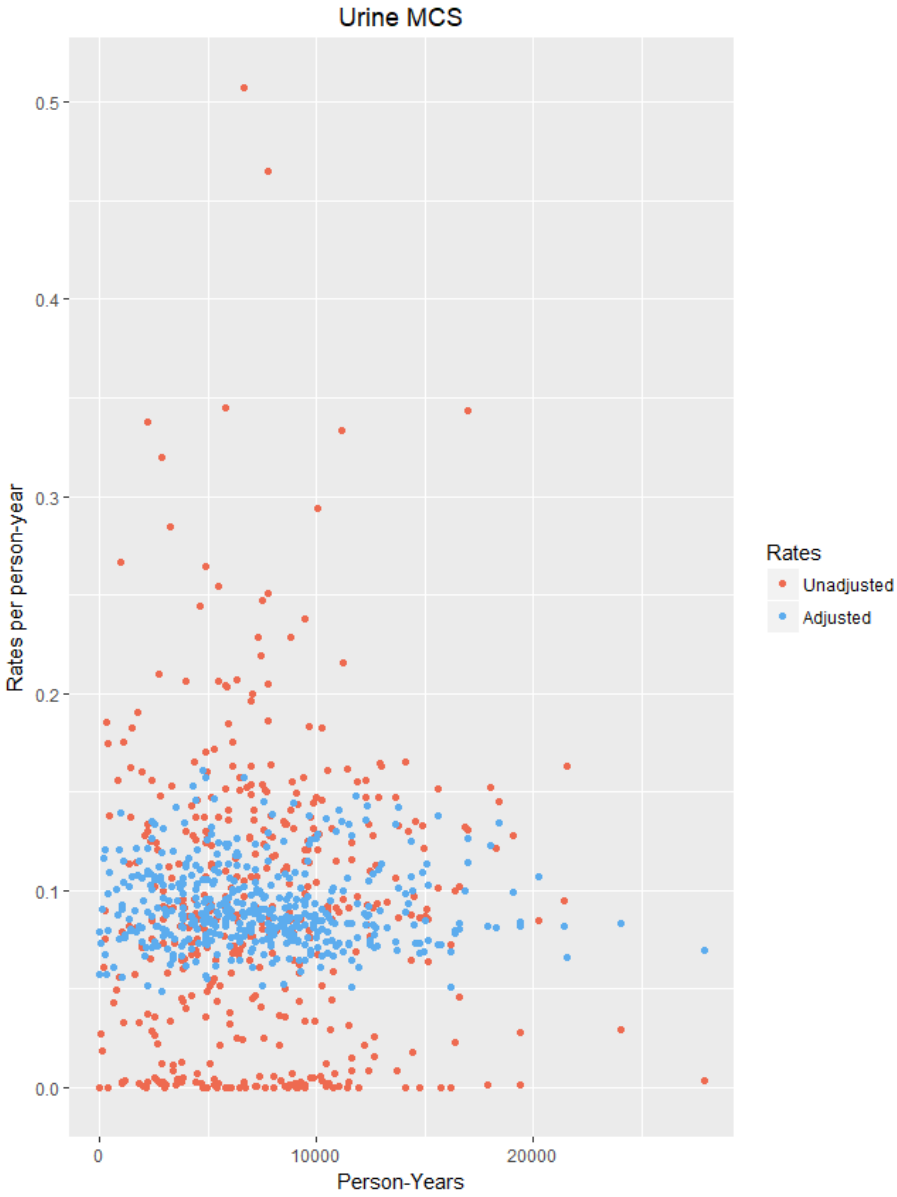

Troponin
0.04
0.03
0.02
0.01
0.00
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

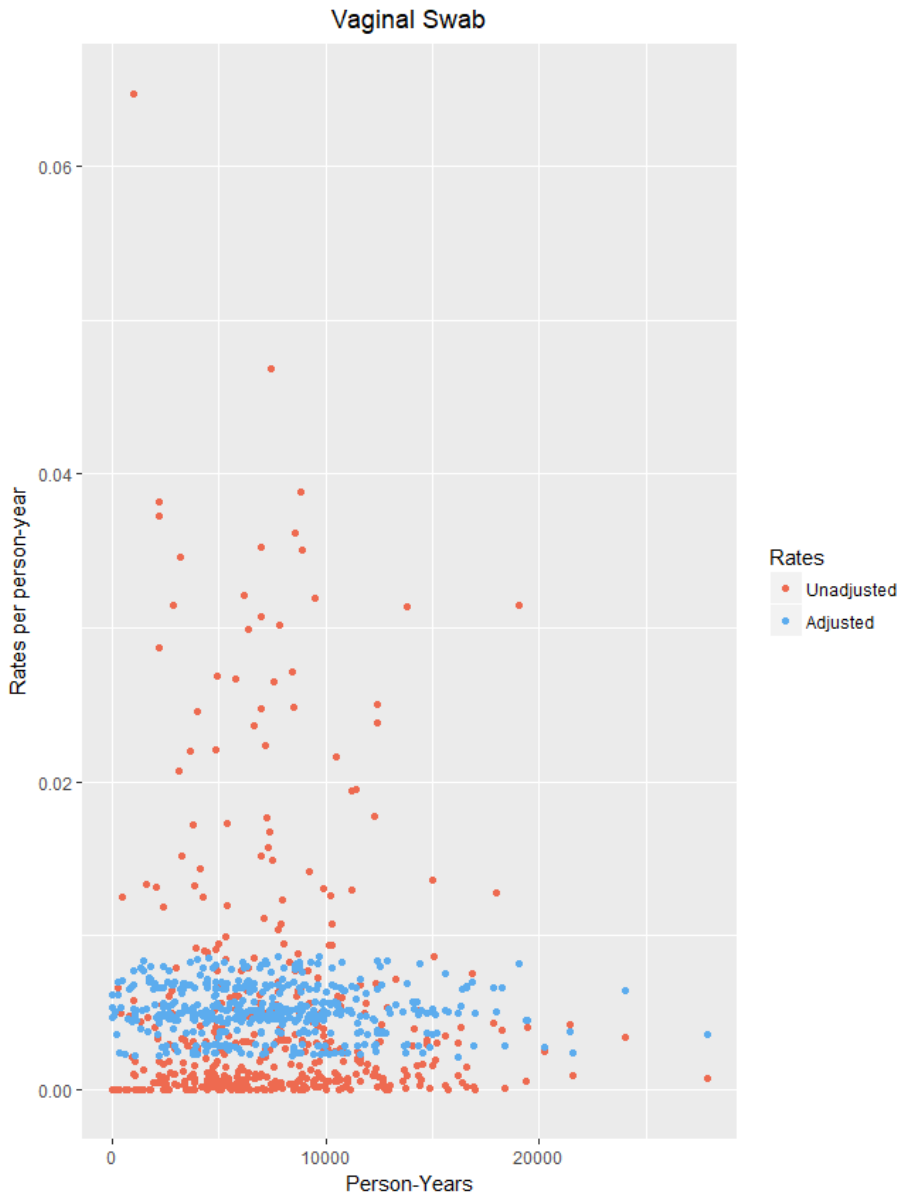

Upper Endoscopy
Rates per person-year
0.05
0.04
0.03
0.02
0.01
0.00
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

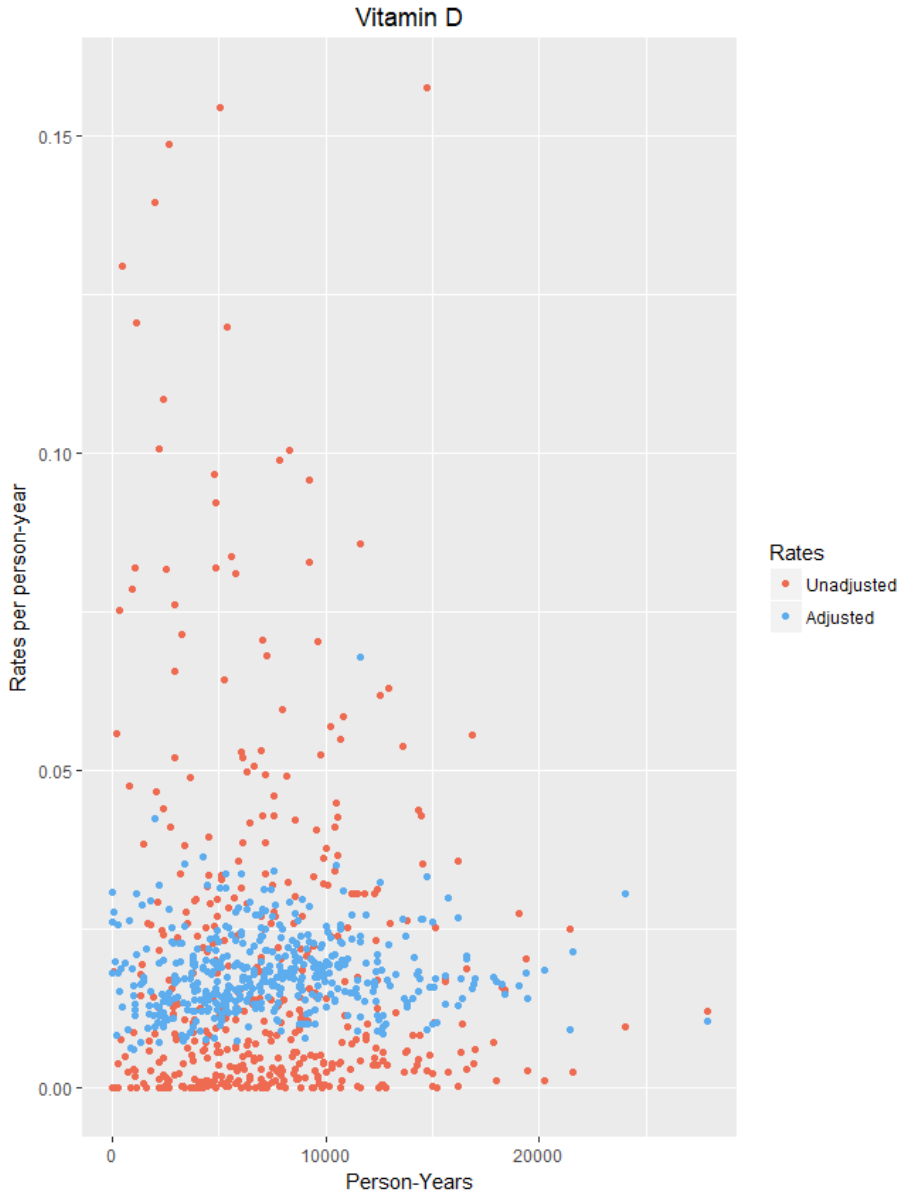

Urine Albumin Creatinine Ratio
Rates per person-year
0.20
0.15
0.10
0.05
0.00
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

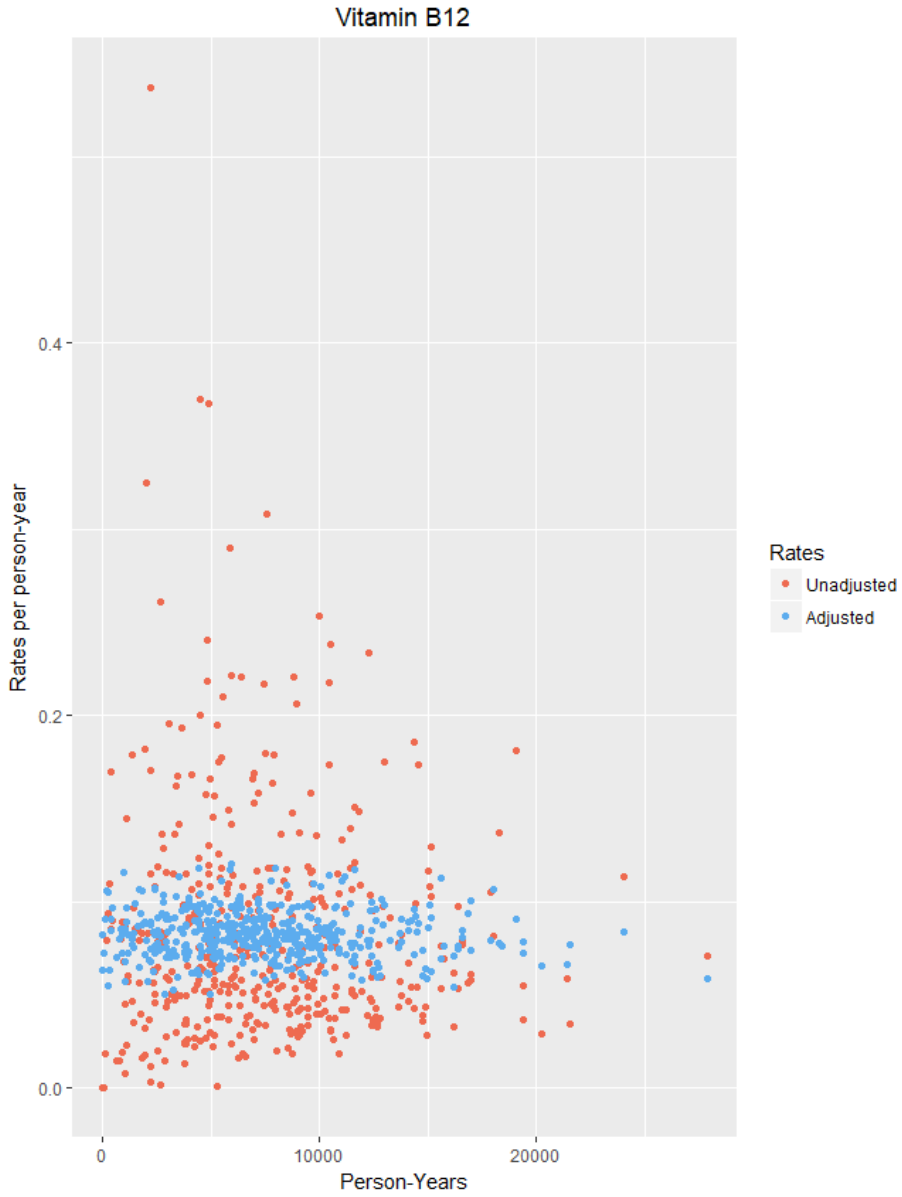

Urine Albumin
0.20
0.15
0.10
0.05
0.00
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

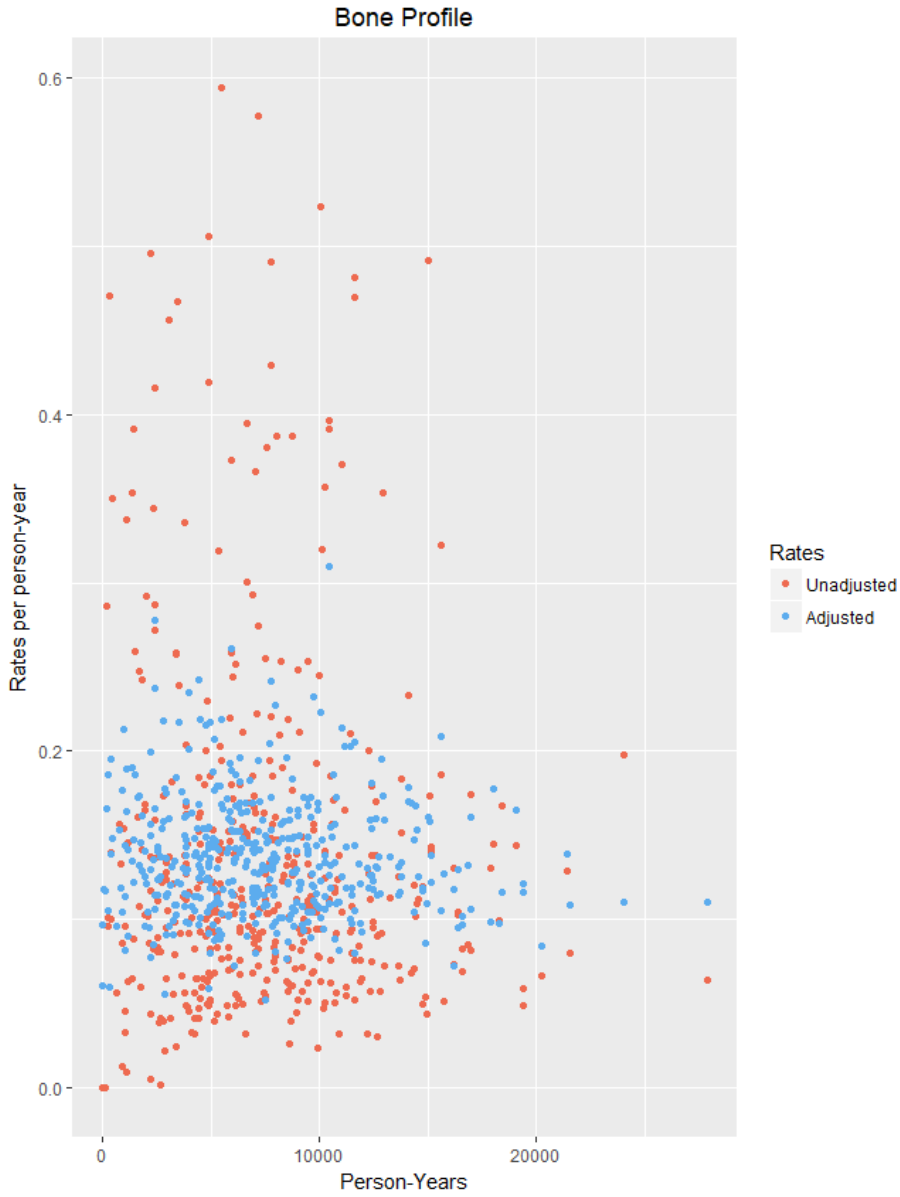

Urine Dipstick
Rates per person-year
0.0
0.1
0.2
0.3
0.4
0.5
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

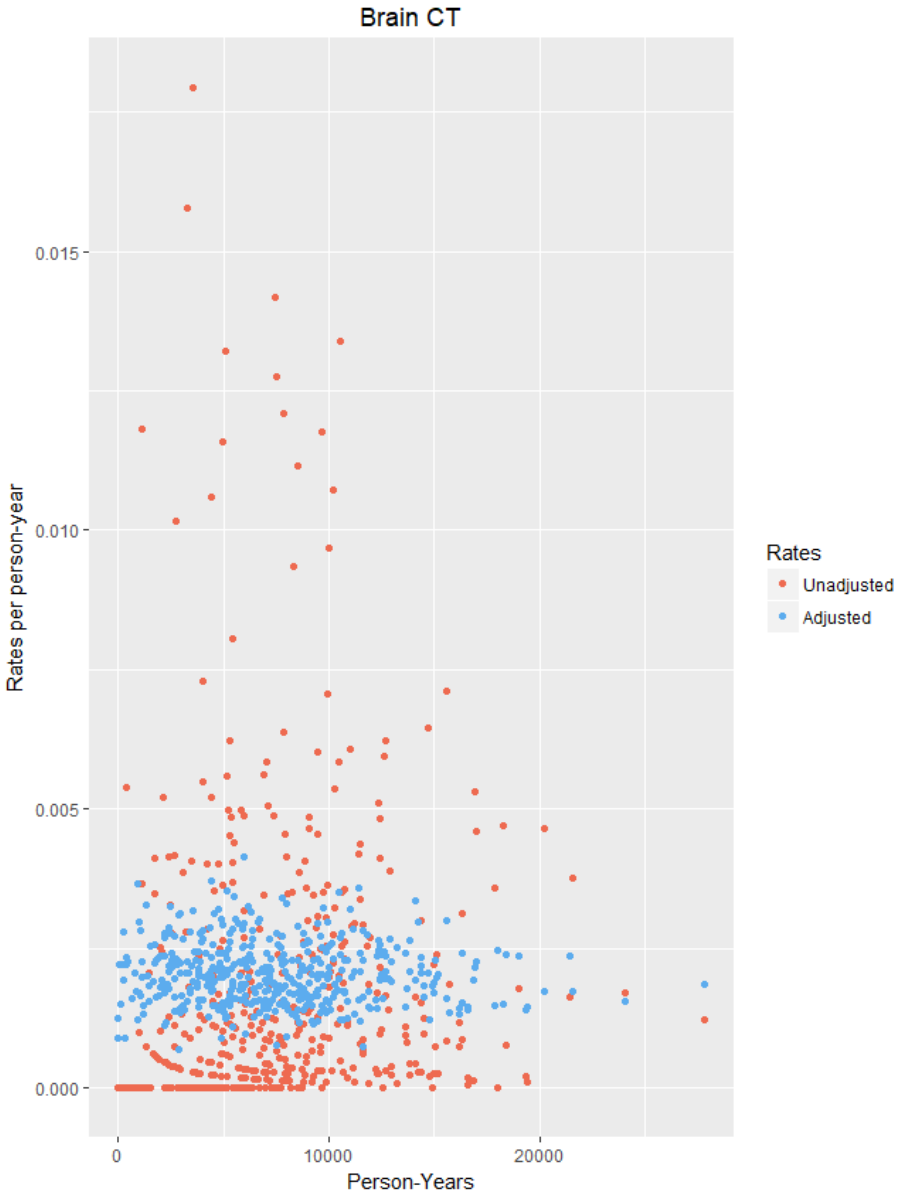

Urine MCS
Rates per person-year
0.5
0.4
0.3
0.2
0.1
0.0
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

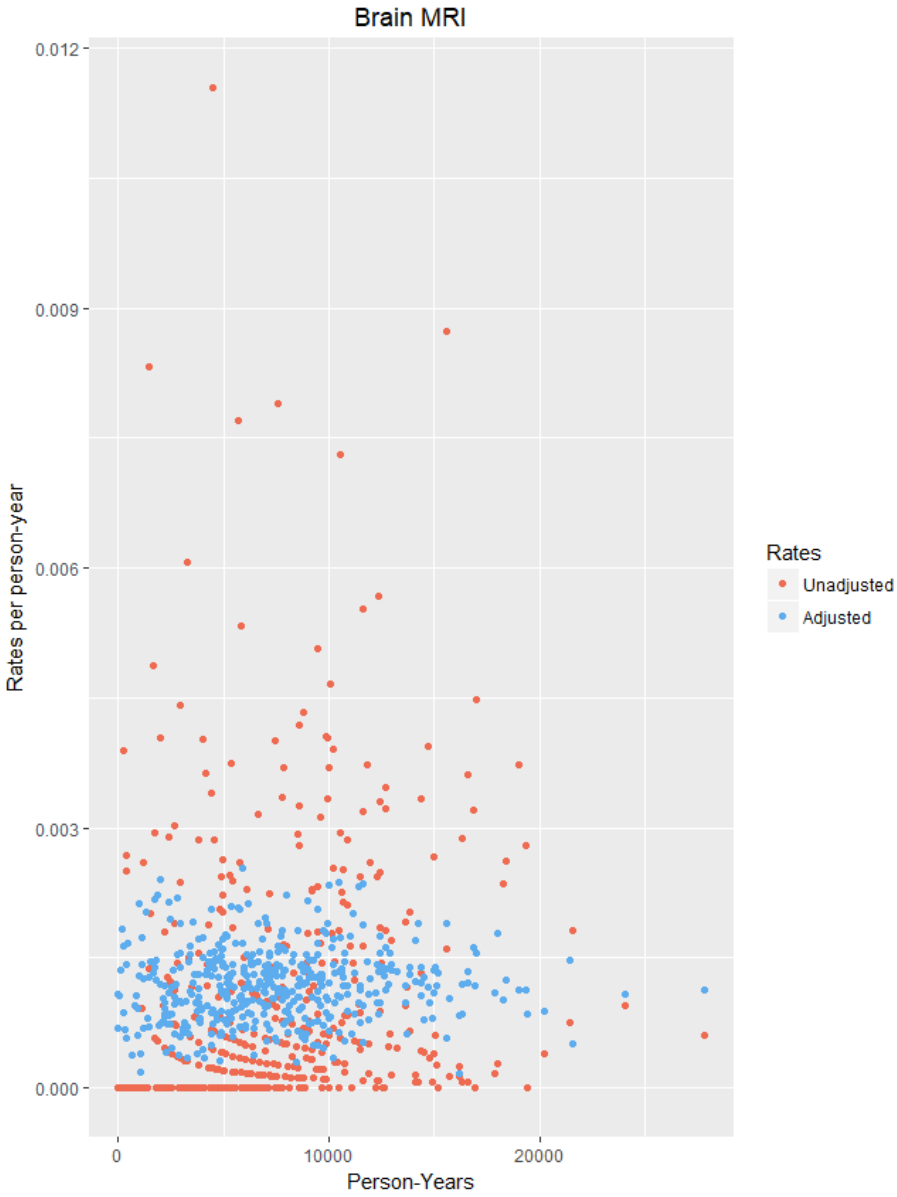

Vaginal Swab
0.06
0.04
0.02
0.00
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted


Vitamin D
0.15
0.10
0.05
0.00
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted


Vitamin B12
Rates per person-year
0.4
0.2
0.0
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

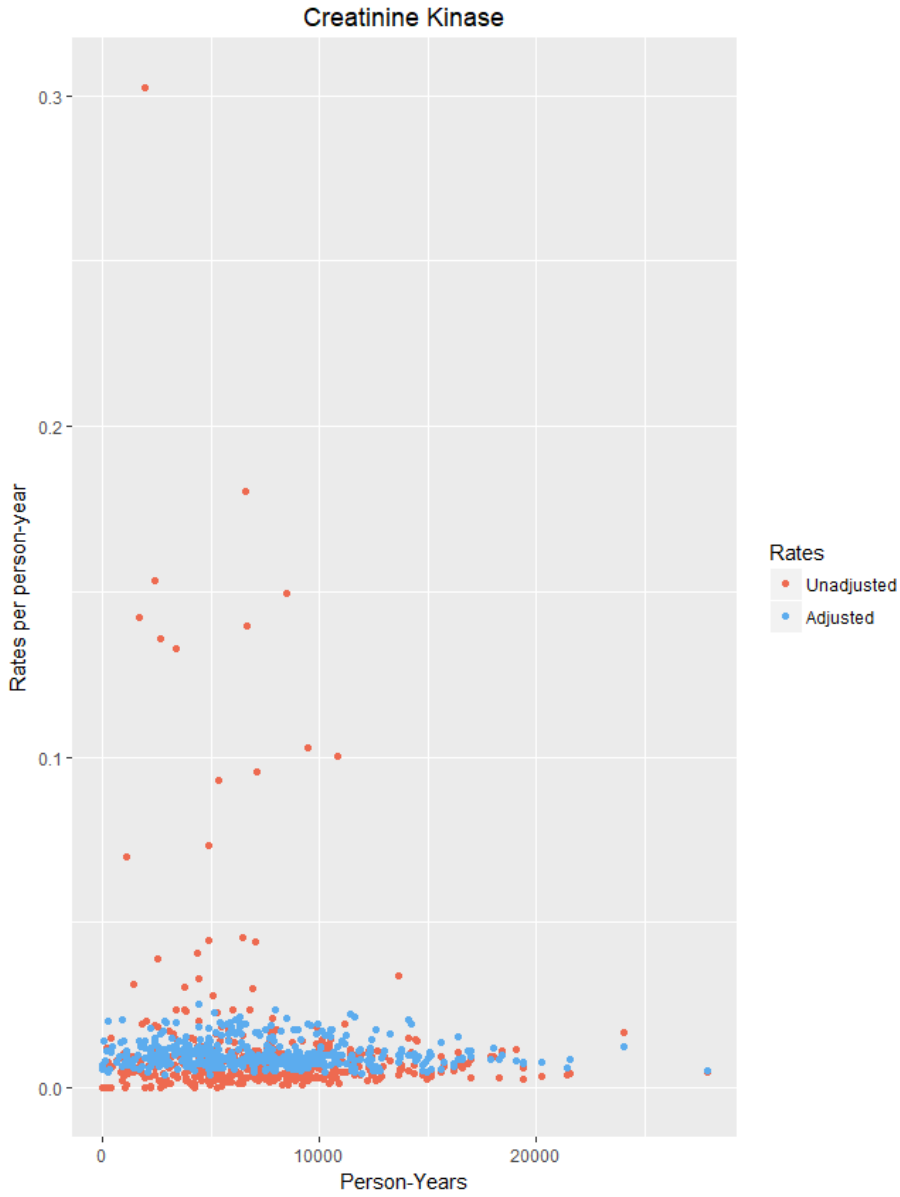

Bone Profile
0.6
0.4
0.2
0.0
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

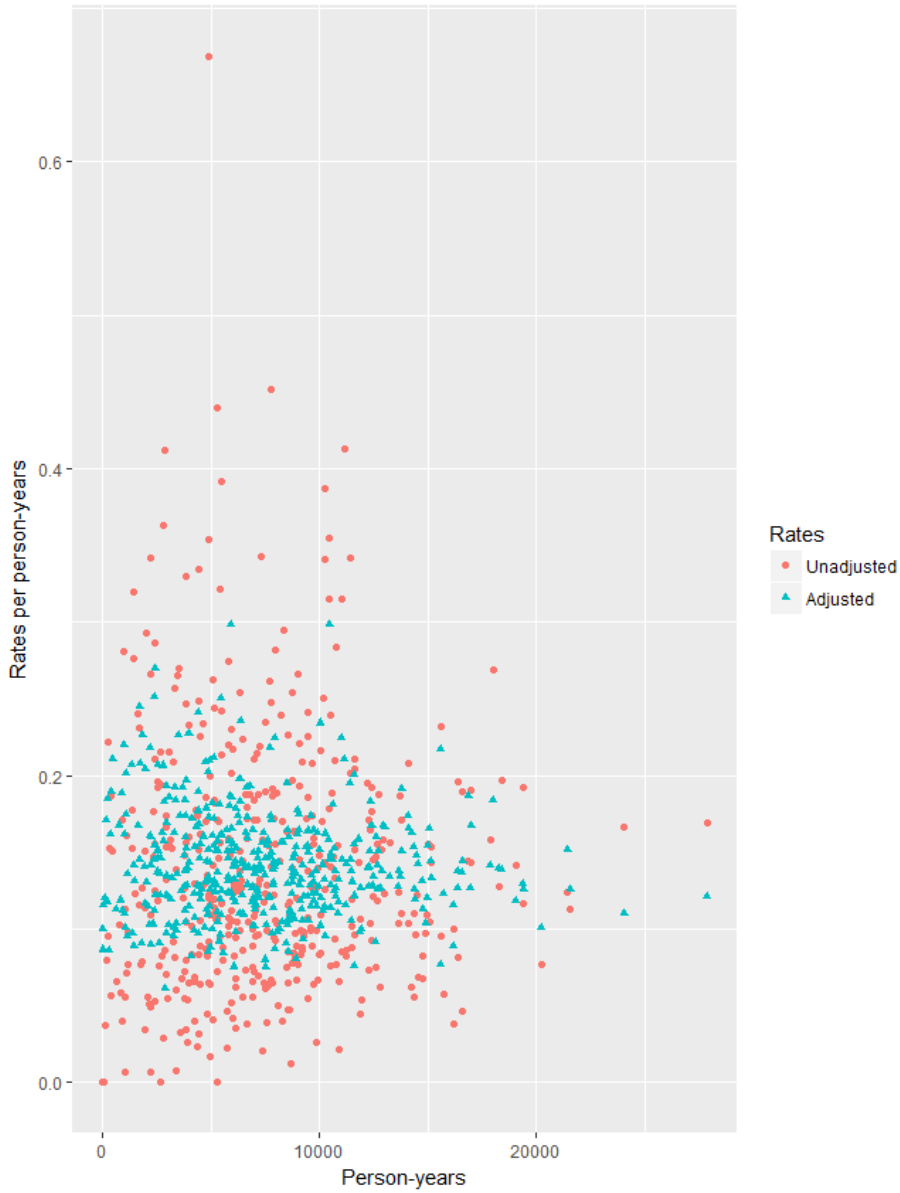

Brain CT
Rates per person-year
0.015
0.010
0.005
0.000
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

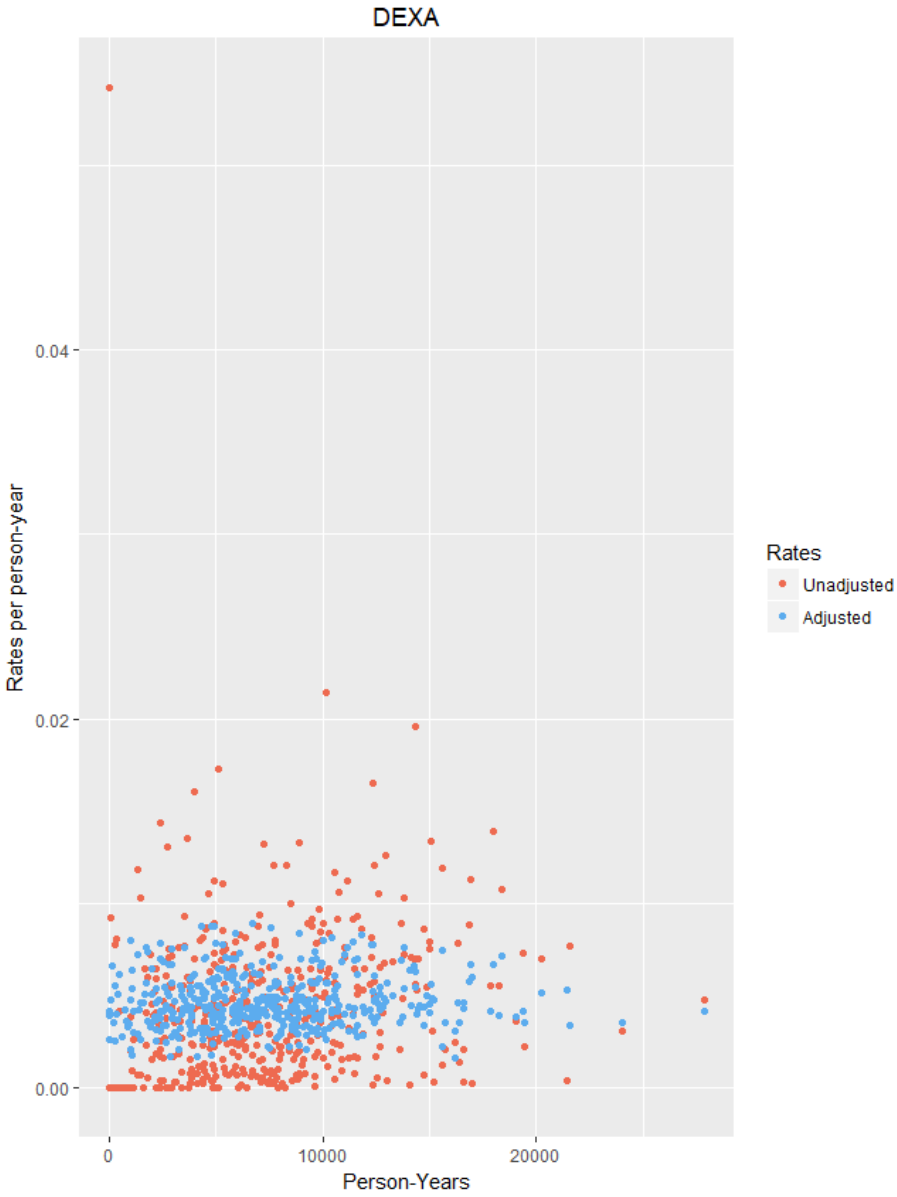

Brain MRI
0.012
0.009
0.006
0.003
0.000
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted


Clotting
0.6
0.4
0.2
0.0
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

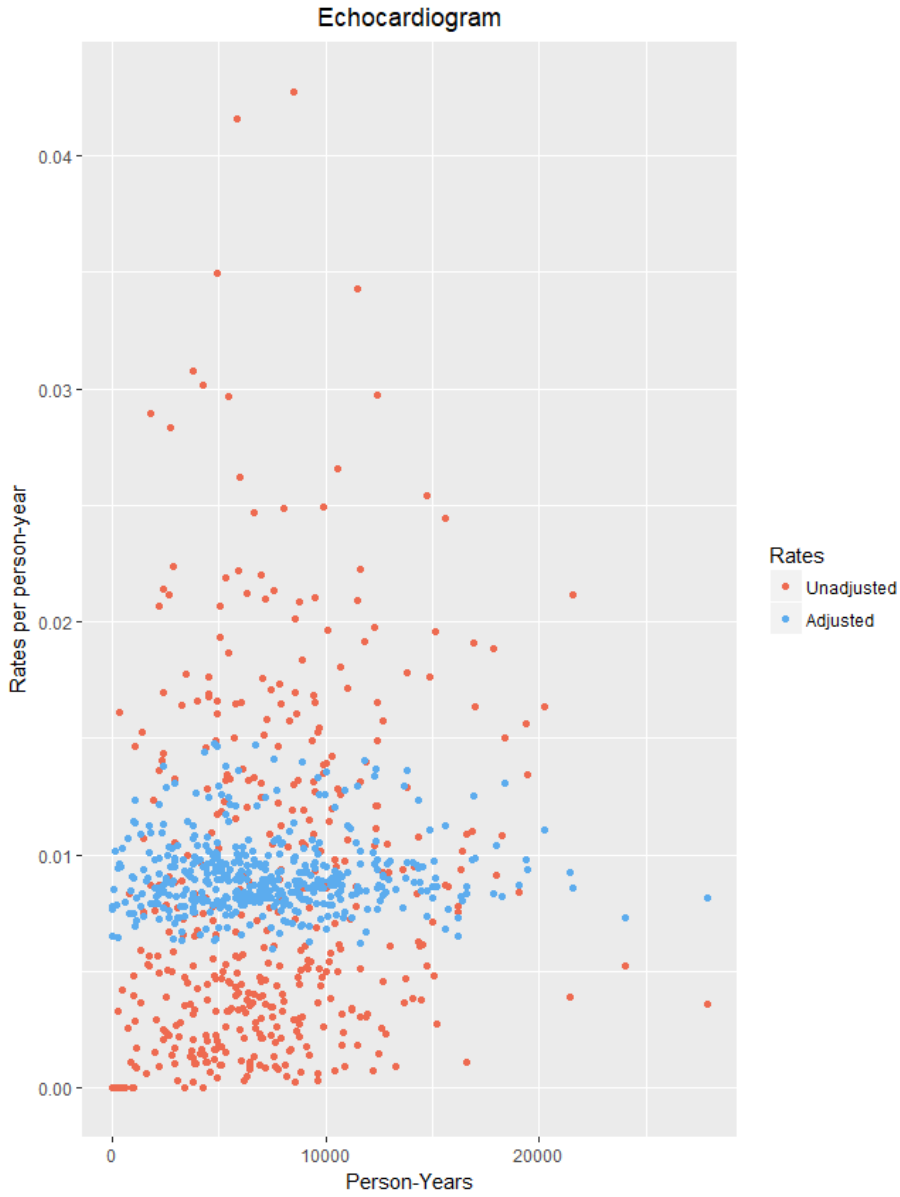

Colonoscopy
0.03
0.02
0.01
0.00
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted


Creatinine Kinase
0.3
0.2
0.1
0.0
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

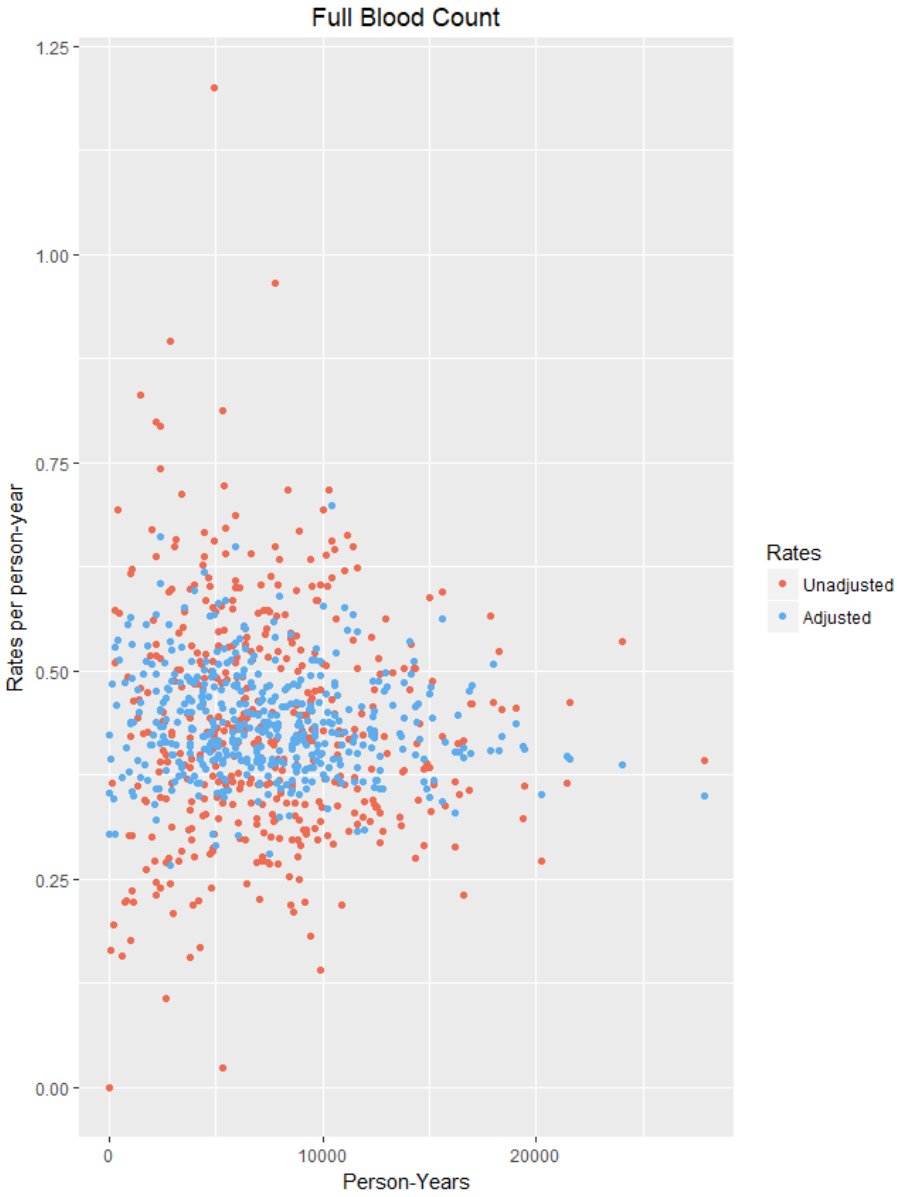

0.6
0.4
0.2
0.0
Rates per person-years
0
10000
20000
Person-years
Rates
Unadjusted
Adjusted

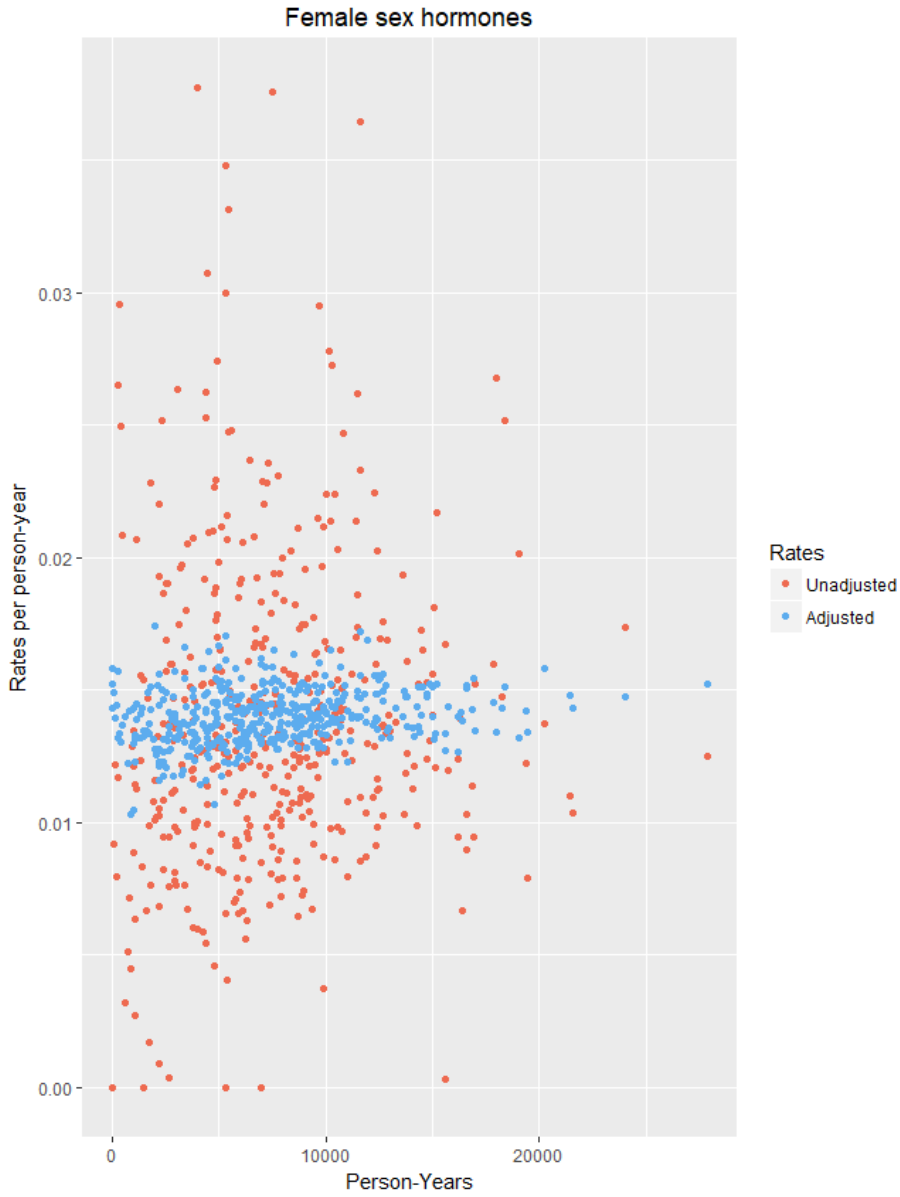

DEXA
Rates per person-year
0.04
0.02
0.00
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

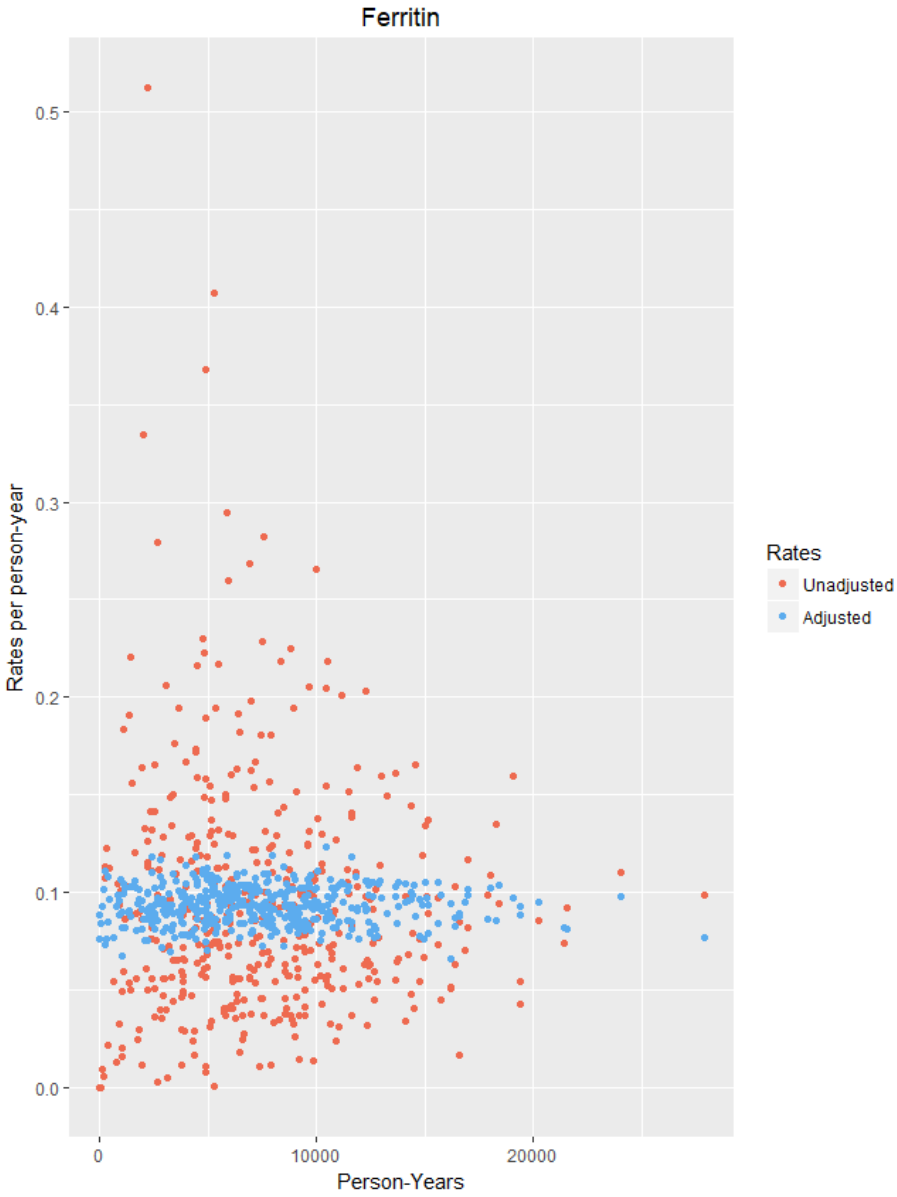

Drug monitoring tests
0.100
0.001
Rates per person-year (logged scale
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

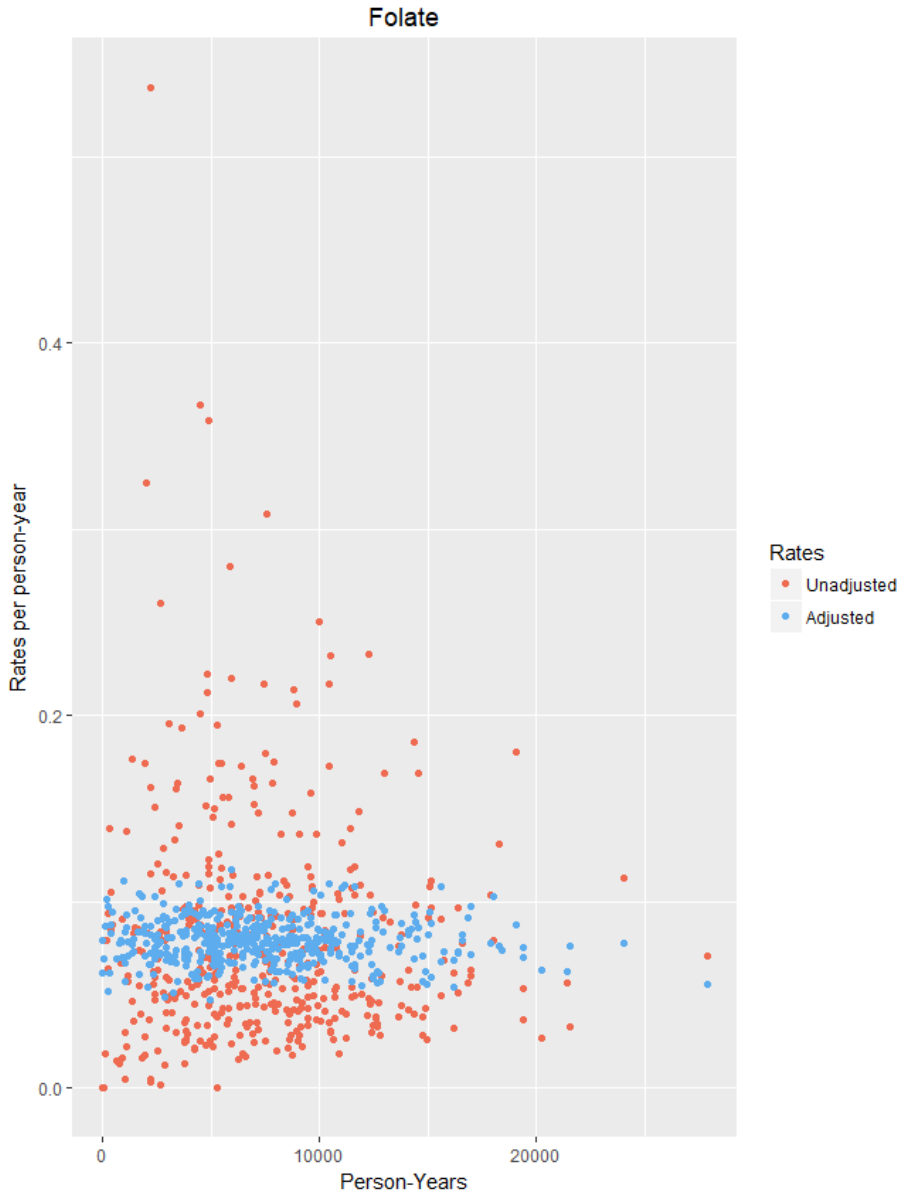

Echocardiogram
0.04
0.03
0.02
0.01
0.00
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

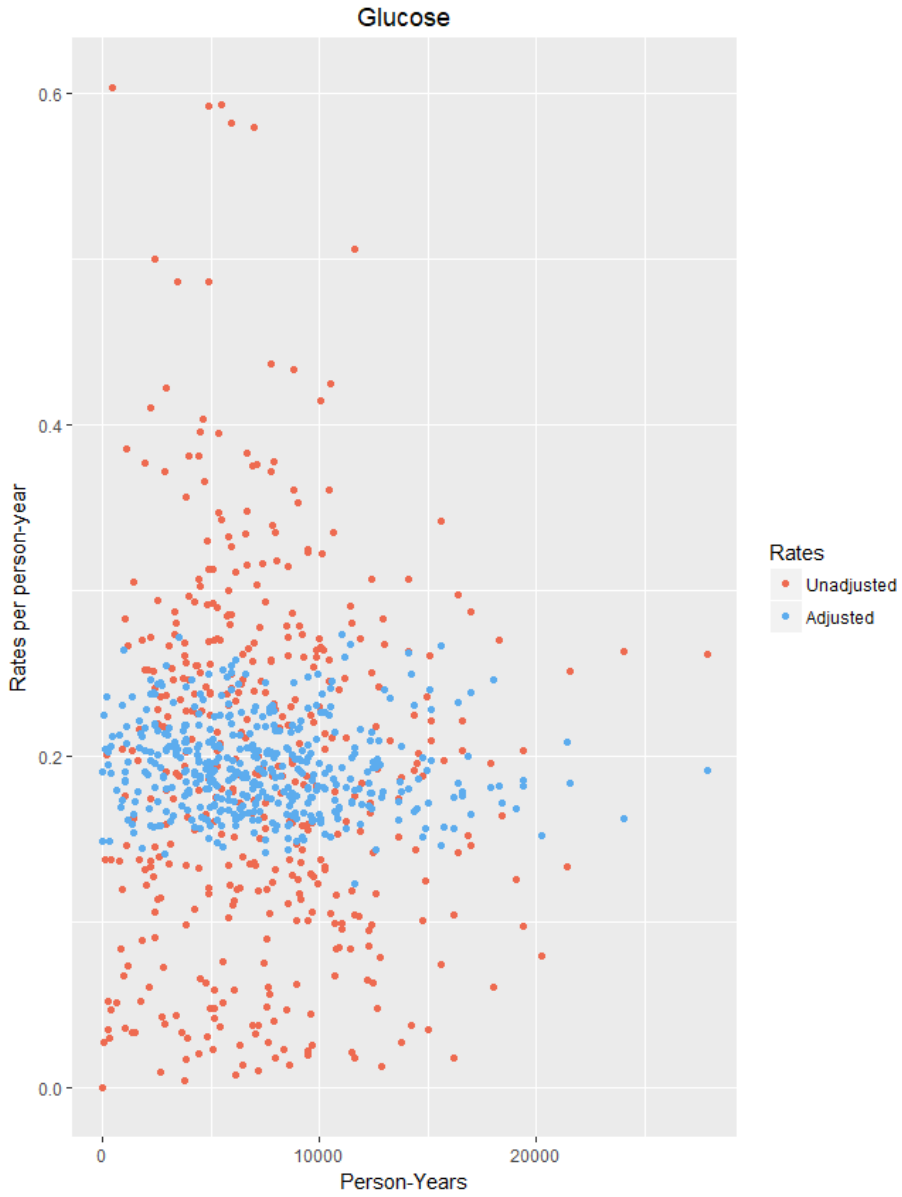

Erythrocyte Sedimentation Rate
Rates per person-year
0.4
0.3
0.2
0.1
0.0
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

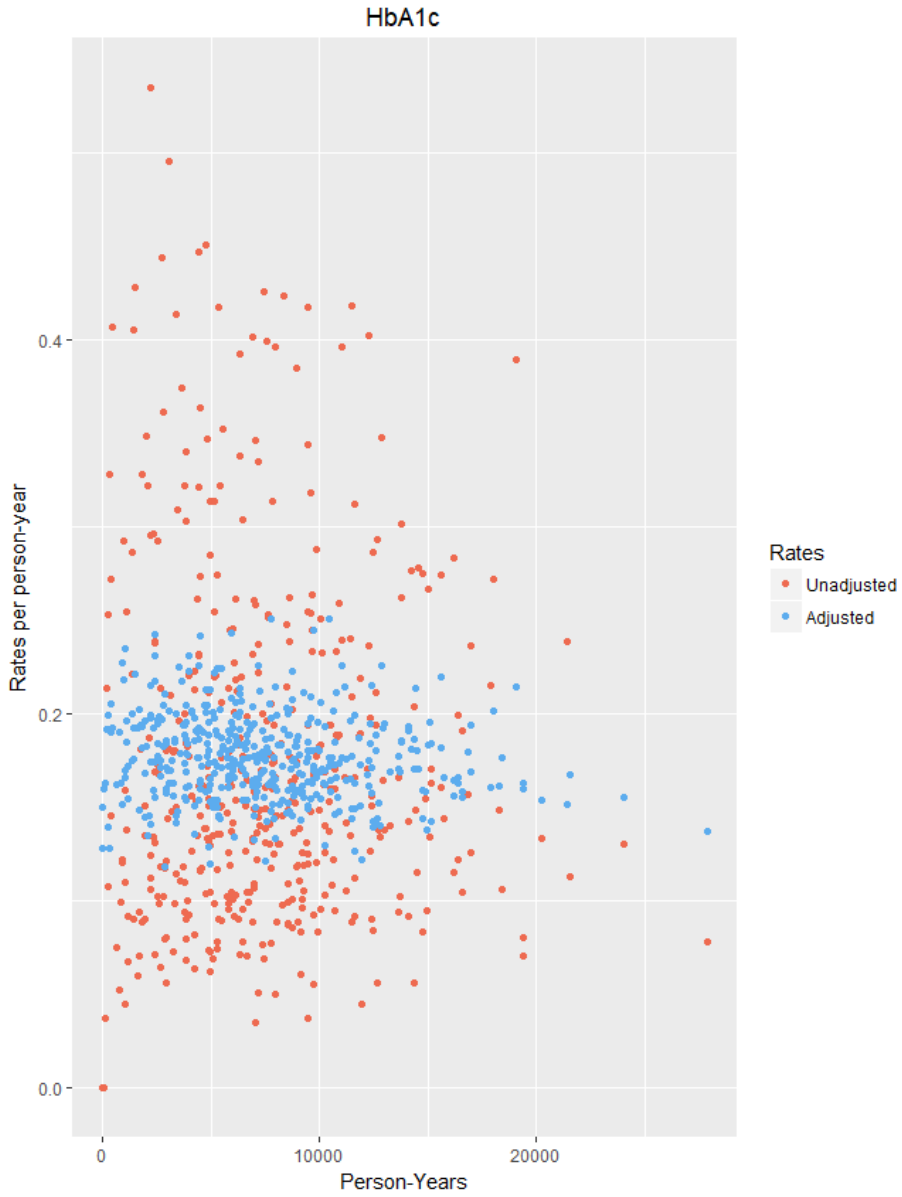

Full Blood Count
1.25
1.00
0.75
0.50
0.25
0.00
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

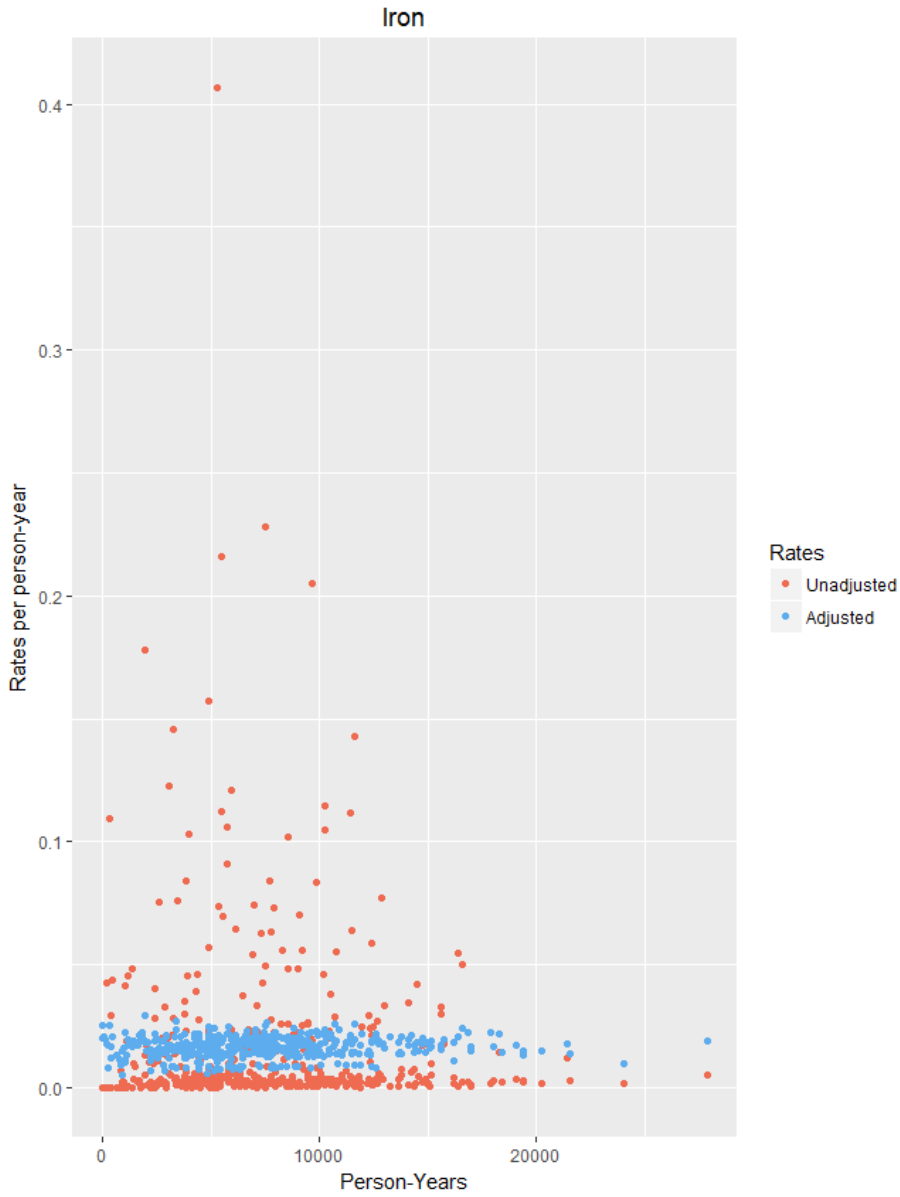

Female sex hormones
Rates per person-year
0.03
0.02
0.01
0.00
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

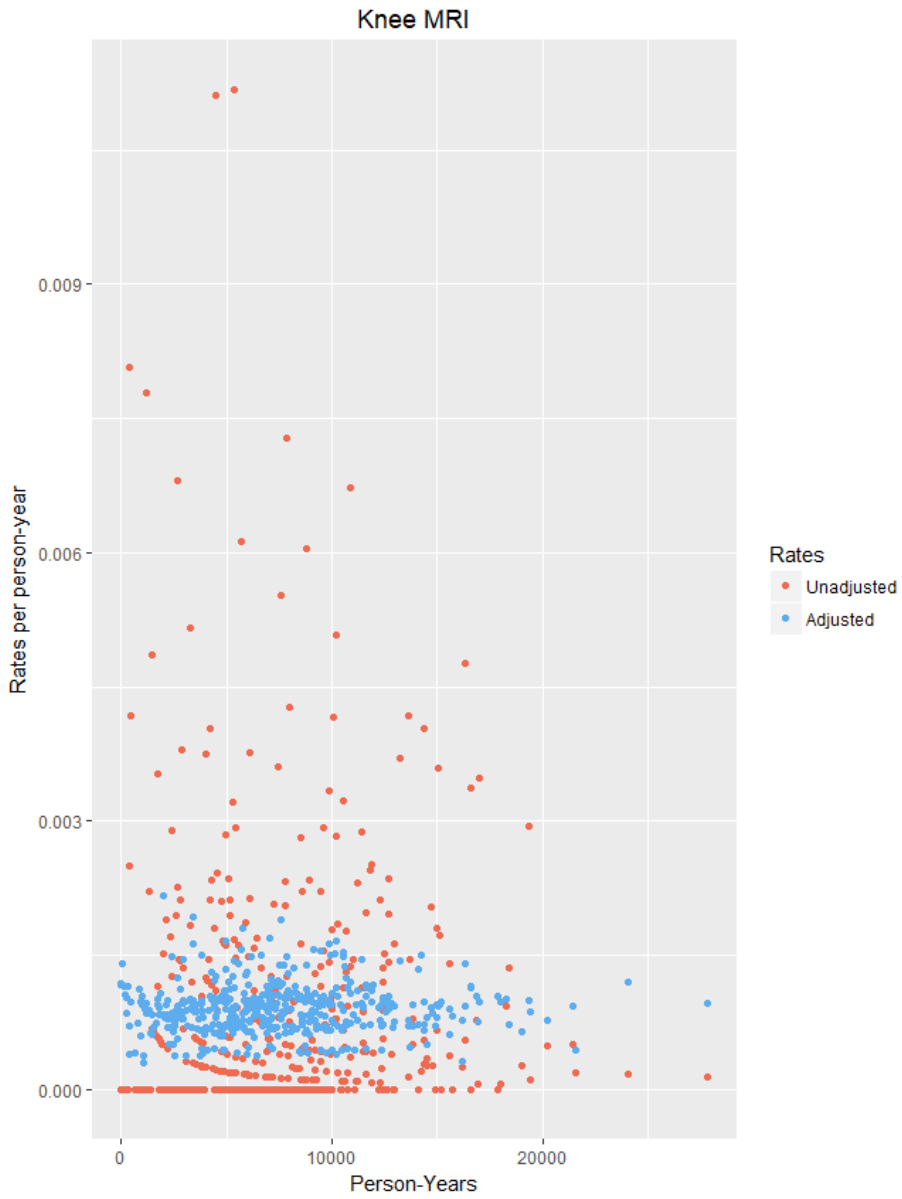

Ferritin
Rates per person-year
0.5
0.4
0.3
0.2
0.1
0.0
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

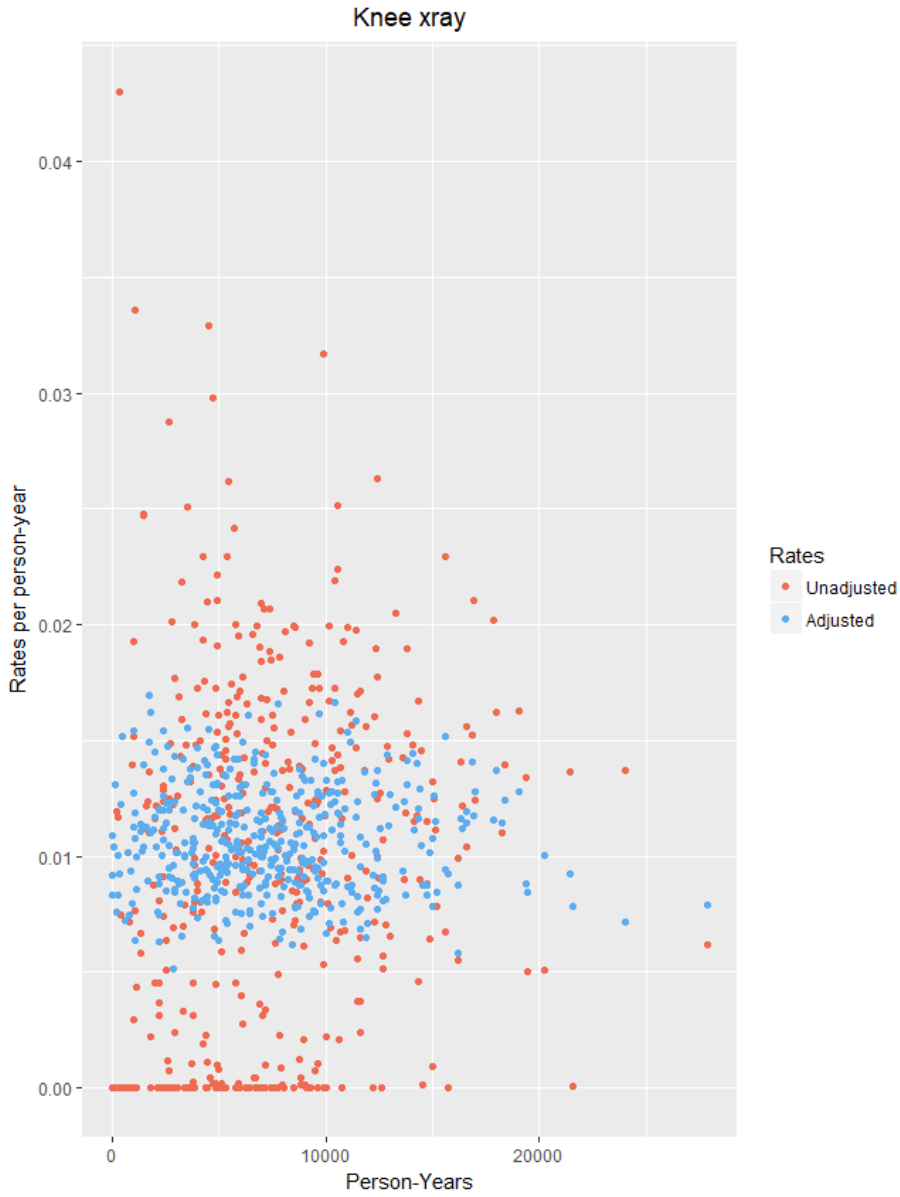

Folate
Rates per person-year
0.4
0.2
0.0
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted


Glucose
0.6
0.4
0.2
0.0
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

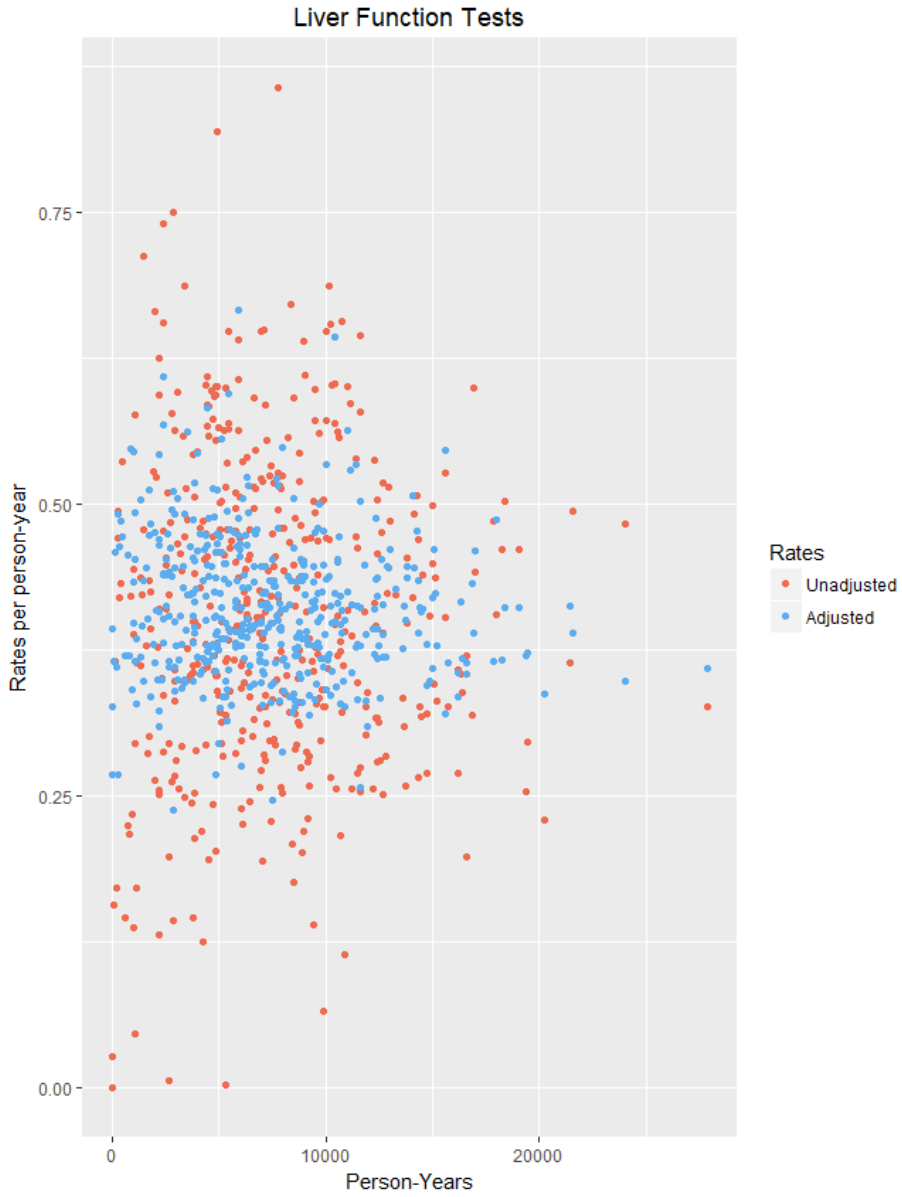

HbA1c
Rates per person-year
0.4
0.2
0.0
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

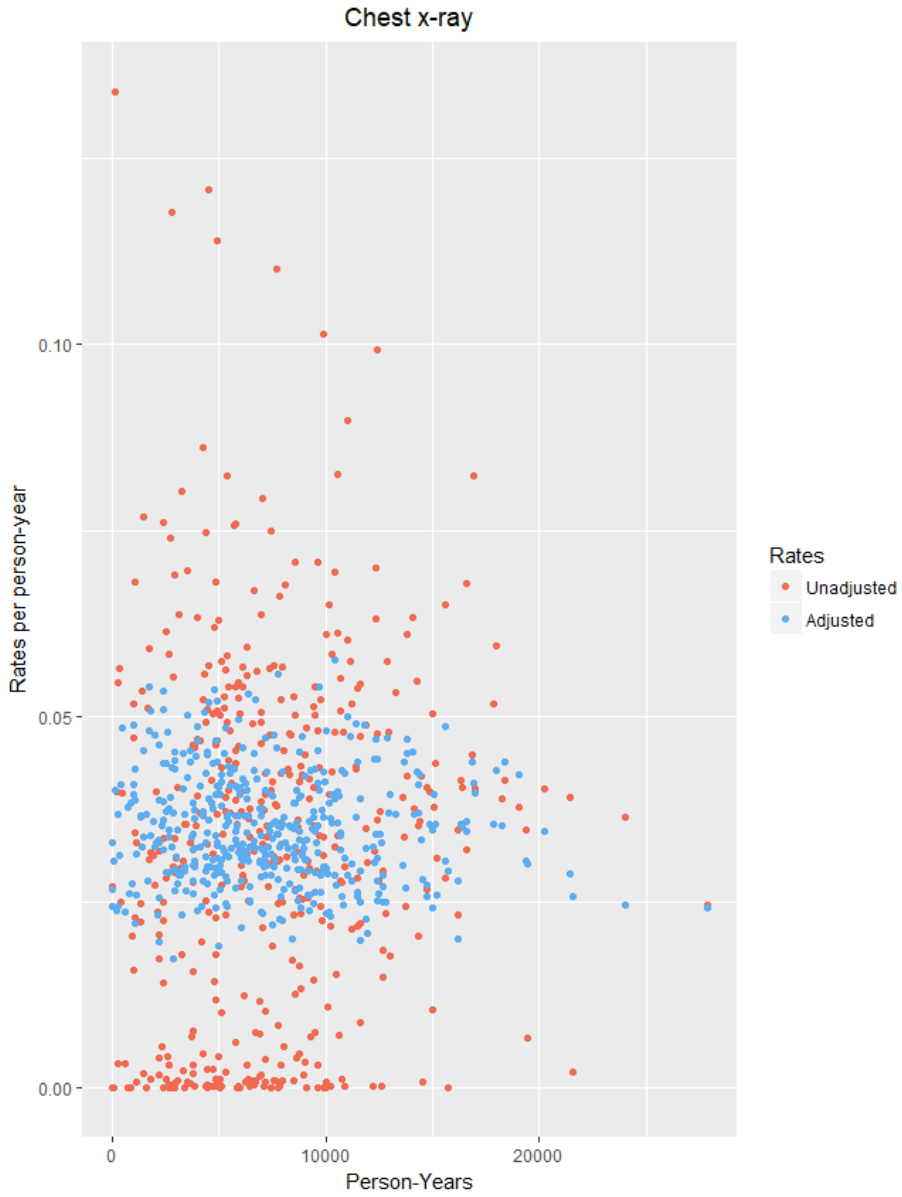

Iron
Rates per person-year
0.4
0.3
0.2
0.1
0.0
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted


Knee MRI
Rates per person-year
0.009
0.006
0.003
0.000
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted


Knee xray
0.04
0.03
0.02
0.01
0.00
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted


Lumbar Spine MRI
Rates per person-year
0.03
0.02
0.01
0.00
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted


Liver Function Tests
Rates per person-year
0.75
0.50
0.25
0.00
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted


Chest x-ray
0.10
0.05
0.00
Rates per person-year
0
10000
20000
Person-Years
Rates
Unadjusted
Adjusted

4.3 RECORD Checklist

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items	Location in manuscript where items are reported
Title and abstract					
	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found		RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included. RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract. RECORD 1.3: If linkage between databases was conducted for the study,	Title and abstract: page 59

				this should be clearly stated in the title or abstract.	
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported			Introduction: Page 59
Objectives	3	State specific objectives, including any prespecified hypotheses			Introduction: Page 59
Methods					
Study Design	4	Present key elements of study design early in the paper			Methods: Page 60 to 62
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection			Methods: Page 60 to 62
Participants	6	*(a) Cohort study* - Give the eligibility criteria, and the sources		RECORD 6.1: The methods of study population selection (such as codes or	Methods: Page 60 to 62

		and methods of selection of participants. Describe methods of follow-up *Case-control study* - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study* - Give the eligibility criteria, and the sources and methods of selection of participants *(b) Cohort study* - For matched studies, give matching criteria and number of exposed and unexposed		algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided. RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with	

		Case-control study - For matched studies, give matching criteria and the number of controls per case		linked data at each stage.	
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.		RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	Methods: Page 60 to 62
Data sources/ measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group			Methods: Page 60 to 62
Bias	9	Describe any efforts to address			Methods: Page 60

		potential sources of bias			to 62
Study size	10	Explain how the study size was arrived at			Methods: Page 60 to 62
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why			Methods: Page 60 to 62
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) *Cohort study* - If applicable, explain how loss to follow-up was			Methods: Page 60 to 62

		addressed *Case-control study* - If applicable, explain how matching of cases and controls was addressed *Cross-sectional study* - If applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses			
Data access and cleaning methods		..		RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population. RECORD 12.2: Authors should provide information on the data cleaning methods	Methods: Page 60 to 62

				used in the study.	
Linkage		..		RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.	Methods: Page 60 to 62
Results					
Participants	13	(a) Report the numbers of individuals at each stage of the study (*e.g.*, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed) (b) Give reasons for non-participation at each stage.		RECORD 13.1: Describe in detail the selection of the persons included in the study (*i.e.,* study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.	Results: Page 63 - 74

		(c) Consider use of a flow diagram			
Descriptive data	14	(a) Give characteristics of study participants (*e.g.*, demographic, clinical, social) and information on exposures and potential confounders (b) Indicate the number of participants with missing data for each variable of interest (c) *Cohort study* - summarise follow-up time (*e.g.*, average and total amount)			Results: Page 63 - 74
Outcome data	15	*Cohort study* - Report numbers of outcome events or summary measures over time *Case-control study* - Report numbers in each exposure			Results: Page 63 - 74

		category, or summary measures of exposure *Cross-sectional study* - Report numbers of outcome events or summary measures			
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorized (c) If relevant, consider translating estimates of relative risk into			Results: Page 63 - 74

		absolute risk for a meaningful time period			
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses			Results: Page 63 – 74 and appendix
Discussion					
Key results	18	Summarise key results with reference to study objectives			Discussion page 75 – 81
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias		RECORD 19.1: Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported.	Discussion page 75 – 81

Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence			Discussion page 75 – 81
Generalisability	21	Discuss the generalisability (external validity) of the study results			Discussion page 75 – 81
Other Information					
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based			Page 60
Accessibility of protocol, raw		..		RECORD 22.1: Authors should provide information on how to access any	

data, and programming code				supplemental information such as the study protocol, raw data, or programming code.	

Appendix Chapter 5

5.1 RECORD checklist

The RECORD statement – checklist of items, extended from the STROBE statement, that should be reported in observational studies using routinely collected health data.

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items	Location in manuscript where items are reported
Title and abstract					
	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was		RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included.	Title and Abstract: **Page 85**

		done and what was found		RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract. RECORD 1.3: If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract.	
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported			**Page 85**
Objectives	3	State specific objectives, including any prespecified hypotheses			**Page 85**
Methods					
Study Design	4	Present key elements of study			**86-88: Methods**

		design early in the paper			
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection			**86-88: Methods**
Participants	6	*(a) Cohort study* - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study* - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study* - Give the		RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed	**86-88: Methods**

		eligibility criteria, and the sources and methods of selection of participants *(b) Cohort study* - For matched studies, give matching criteria and number of exposed and unexposed *Case-control study* - For matched studies, give matching criteria and the number of controls per case		methods and results should be provided. RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.	
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.		RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	**86-88: Methods** GitHub link
Data sources/	8	For each variable of interest, give			**86-88: Methods**

measurement		sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group			
Bias	9	Describe any efforts to address potential sources of bias			**86-88: Methods**
Study size	10	Explain how the study size was arrived at			Not applicable
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why			**86-88: Methods**
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding			**86-88: Methods**

		(b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) *Cohort study* - If applicable, explain how loss to follow-up was addressed *Case-control study* - If applicable, explain how matching of cases and controls was addressed *Cross-sectional study* - If applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses			

Data access and cleaning methods		..		RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population. RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.	**86-88: Methods**
Linkage		..		RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.	Not applicable
Results					
Participants	13	(a) Report the numbers of individuals at each stage of the		RECORD 13.1: Describe in detail the selection of the persons included in the	**Page 85-92: Results**

		study (*e.g.*, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed) (b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram		study (*i.e.,* study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.	
Descriptive data	14	(a) Give characteristics of study participants (*e.g.*, demographic, clinical, social) and information on exposures and potential confounders (b) Indicate the number of participants with missing data for each variable of interest (c) *Cohort study* - summarise			**Page 85-92:** **Results**

		follow-up time (*e.g.*, average and total amount)			
Outcome data	15	*Cohort study* - Report numbers of outcome events or summary measures over time *Case-control study* - Report numbers in each exposure category, or summary measures of exposure *Cross-sectional study* - Report numbers of outcome events or summary measures			
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were			**Page 85-92: Results**

		adjusted for and why they were included (b) Report category boundaries when continuous variables were categorized (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period			
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses			**Page 85-92: Results**
Discussion					
Key results	18	Summarise key results with reference to study objectives			**Page 92 -96: Discussion**
Limitations	19	Discuss limitations of the study,		RECORD 19.1: Discuss the implications	**Page 92 -96:**

		taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias		of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported.	**Discussion**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence			**Page 92 -96: Discussion**
Generalisability	21	Discuss the generalisability (external validity) of the study results			**Page 92 -96: Discussion**
Other Information					

Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based			**Page 92 -96:**
Accessibility of protocol, raw data, and programming code		..		RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	**Page 92 -96:** **Discussion**

*Reference: Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM, the RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Medicine* 2015; in press.

5.2 The unadjusted and adjusted average proportion of abnormal CRP and ESR tests and total tests ordered per practice

Practice ID	Adjusted average proportion of abnormal ESR tests	Unadjusted average proportion of abnormal ESR tests	Total ESR tests ordered	Adjusted average proportion of abnormal CRP tests	Unadjusted average proportion of abnormal CRP tests	Total CRP tests ordered
1	0.249098	0.25	784	0.262037	0.255783	2248
2	0.223219	0.229508	122	0.363702	0.369461	1670
3	0.153799	0.096154	52	0.261058	0.252269	1102
4	0.165622	0.155	400	0.249819	0.245474	1381
5	0.19622	0.187599	629	0.254637	0.242448	1258
6	0.181387	0.170732	41	0.224107	0.219536	1034
7	0.223473	0.202128	94	0.314022	0.322679	657
8	0.18626	0.184211	152	0.322787	0.323409	2452
9	0.191368	0.152672	131	0.318597	0.344353	363
10	0.202515	0.203008	133	0.326114	0.326269	2265
11	0.228991	0.23219	379	0.374675	0.372024	2352
12	0.207168	0.193182	88	0.332636	0.333025	1081
13	0.236322	0.222222	54	0.422029	0.425871	1693

14	0.120509	0.097674	215	0.331811	0.336764	1749
15	0.200024	0.196646	656	0.347439	0.34621	2071
16	0.227228	0.225933	509	0.33269	0.333061	3669
17	0.192391	0.186761	423	0.319374	0.316727	2750
18	0.185493	0.181818	396	0.325242	0.321462	1067
19	0.231789	0.227907	430	0.326767	0.326894	1967
20	0.24669	0.253968	189	0.307597	0.305925	3004
21	0.206154	0.189189	74	0.343057	0.355072	690
22	0.22705	0.228972	428	0.334889	0.332793	1851
23	0.195901	0.193732	351	0.382956	0.39426	662
24	0.190941	0.189802	1059	0.33091	0.333008	4099
25	0.274649	0.274436	532	0.345126	0.345276	3461
26	0.255595	0.253472	864	0.329283	0.330504	2003
27	0.27942	0.361111	36	0.320793	0.330606	611
28	0.151506	0.142857	455	0.283214	0.280108	2956
29	0.314102	0.320132	303	0.341384	0.341006	1868
30	0.20582	0.194286	175	0.307482	0.304534	1632
31	0.171587	0.154639	97	0.302994	0.300578	2076
32	0.245864	0.280702	57	0.22189	0.211034	725

33	0.22394	0.223881	469	0.33931	0.340122	3290
34	0.247565	0.253289	608	0.311518	0.313179	2944
35	0.19641	0.220339	59	0.280564	0.28113	1558
36	0.169774	0.164645	577	0.349038	0.349145	3099
37	0.15263	0.071429	14	0.390376	0.431373	153
38	0.238568	0.238757	1512	0.328851	0.329273	5090
39	0.223604	0.220657	639	0.328627	0.332777	2395
40	0.238758	0.240275	437	0.332329	0.332856	2794
41	0.243861	0.244565	552	0.327481	0.328089	2606
42	0.204713	0.202814	853	0.345554	0.350959	1929
43	0.211888	0.212389	565	0.344276	0.347458	2596
44	0.226802	0.228856	201	0.324132	0.315907	899
45	0.201822	0.2	375	0.290282	0.287429	1583
46	0.199989	0.188498	313	0.328451	0.32596	1718
47	0.157064	0.149635	822	0.318591	0.323546	1771
48	0.19624	0.191885	1183	0.336693	0.338617	3470
49	0.192618	0.189373	734	0.347636	0.349663	3266
50	0.21993	0.218527	421	0.328321	0.327919	2278
51	0.233643	0.236111	72	0.342967	0.343793	2167

52	0.228171	0.244898	49	0.368238	0.372881	1829
53	0.283076	0.293907	279	0.265054	0.253623	828
54	0.320004	0.334601	263	0.352685	0.350658	1520
55	0.334528	0.33665	603	0.312891	0.316848	1193
56	0.380329	0.383205	1691	0.330093	0.330702	2534
57	0.355271	0.385714	140	0.295993	0.287866	3519
58	0.355076	0.374517	259	0.333127	0.333646	2134
59	0.262564	0.311111	90	0.341991	0.345125	2318
60	0.327801	0.332378	1047	0.357451	0.359545	4222
61	0.212604	0.213615	426	0.31787	0.317635	2994
62	0.343735	0.357934	271	0.239154	0.221068	674
63	0.301674	0.320755	106	0.300869	0.288136	472
64	0.323381	0.32504	1889	0.266461	0.262742	3924
65	0.348004	0.351794	1282	0.342103	0.342095	2625
66	0.404857	0.413953	645	0.364423	0.363771	3373
67	0.35122	0.358811	471	0.393269	0.397295	1996
68	0.28539	0.302198	182	0.326412	0.330532	1071
69	0.221281	0.219373	351	0.359071	0.362224	2446

Appendix Chapter 6

6.1 PRISM Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Page 99
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	In paper not thesis
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Page 99
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	Page 100
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Page 100
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Page 100/1
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Page 100
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Page 100 - 103 & appendix
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Page 100 – 103
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Page 100 – 103

Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	Page 100 - 103 & appendix
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	Page 100 - 103
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	Page 100 - 103

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	Page 100 - 103
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	Page 100 - 103
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Page 103- 109
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Page 103- 109
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Page 103- 109
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Page 103- 109
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Page 103- 109
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Page 103- 109
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	Page 103- 109
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	Page 109 - 114
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	Page 109 -

			114
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Page 109 - 114
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	NA

6.2 Search strategy

MEDLINE Search Strategy

1. Ambulatory Care/

2. exp Ambulatory Care Facilities/

3. general practice/ or family practice/

4. general practitioners/ or physicians, family/ or physicians, primary care/

5. Primary Health Care/

6. Office visits/

7. Academic Medical Centers/

8. (ambulatory adj3 (care or setting? or facilit* or ward? or department? or service?)).ti,ab.

9. ((general or family) adj2 (practi* or physician? or doctor?)).ti,ab.

10. (primary care or primary health care or primary healthcare or family medicine or community medicine or community health).ti,ab.

11. (gp or gps).ti,ab.

12. (after hour? or afterhour? or "out of hour?" or ooh).ti,ab.

13. (clinic? or visit?).ti,ab.

14. ((health* or medical) adj2 (center? or centre?)).ti,ab.

15. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14

16. exp Emergency Service, Hospital/

17. Emergency Medical Services/

18. (emergency adj3 (care or setting? or facilit* or ward? or department? or service? or room?)).ti,ab.

19. (emergency medicine or ed or er or a&e).ti,ab.

20. 16 or 17 or 18 or 19

21. 15 or 20

22. guidelines as topic/ or practice guidelines as topic/

23. (guideline? or guidance?).ti,ab.

24. 22 or 23

25. (adhere* or non-adhere* or nonadhere* or concord* or non-concord* or nonconcord* or discord* or comply or complian* or non-complian* or noncomplian* or align* or nonalign* or nonalign* or congruen* or incongruen* or consisten* or inconsisten* or contradict*).ti,ab.

26. ((does or "does not" or doesn?t or did or "did not" or didn?t or "not" or fail*) adj3 (follow* or met or meet or meeting or match or matching or "in line with")).ti,ab.

27. ((follow* or met or meet or meeting or match or matching or "in line with" or keep or kept or keeping or utili?ation or utile?e? or change?) adj5 (criteria or recommend* or guideline? or guidance)).ti,ab.

28. Physician's Practice Patterns/

29. clinical competence/ or nursing competence/

30. 25 or 26 or 27 or 28 or 29

31. 24 and 30

32. Guideline Adherence/

33. 31 or 32

34. exp *"diagnostic techniques and procedures"/

35. exp "diagnostic techniques and procedures"/ut

36. (diagnos* or detect* or test* or screen* or manag*).ti.

37. (imaging or radiolog* or tomogra* or ct scan* or pet scan* or echocardiogra* or angiogra* or ultrasound* or ultrasonogra*).ti,ab.

38. ((medical or clinical or diagnos* or screening or routine or laboratory) adj5 (test* or investigation?)).ti,ab.

39. ((h?ematolog* or blood or urin* or saliva*) adj5 test*).ti,ab.

40. ((stress* or physical or function*) adj5 test*).ti,ab.

41. 34 or 35 or 36 or 37 or 38 or 39 or 40

42. 21 and 33 and 41

43. ((necessary or unnecessary or appropriate* or inappropriate* or waste* or utili?ation or indicated or excess* or less or more or increas* or decreas*) adj10 (test* or imaging or radiolog* or tomogra* or ct scan* or pet scan* or echocardiogra* or angiogra* or ultrasound* or ultrasonogra* or investigation?)).ti,ab.

44. ((order* or request*) adj5 (test* or imaging or radiolog* or tomogra* or ct scan* or pet scan* or echocardiogra* or angiogra* or ultrasound* or ultrasonogra* or investigation?)).ti,ab.

45. Unnecessary Procedures/

46. 43 or 44 or 45

47. 21 and 24 and 46

48. 21 and 41 and 45

49. 42 or 47 or 48

50. limit 49 to yr="1999 -Current"

51. limit 50 to english language

52. exp animals/ not humans.sh.

53. 51 not 52

EMBASE Search Strategy

1. Ambulatory Care/

2. general practice/

3. general practitioners/

4. Primary Health Care/

5. (ambulatory adj3 (care or setting? or facilit* or ward? or department? or service?)).ti,ab.

6. ((general or family) adj2 (practi* or physician? or doctor?)).ti,ab.

7. (primary care or primary health care or primary healthcare or family medicine or community medicine or community health).ti,ab.

8. (gp or gps).ti,ab.

9. (after hour? or afterhour? or "out of hour?" or ooh).ti,ab.

10. (clinic? or visit?).ti,ab.

11. ((health* or medical) adj2 (center? or centre?)).ti,ab.

12. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11

13. Emergency Ward/

14. (emergency adj3 (care or setting? or facilit* or ward? or department? or service? or room?)).ti,ab.

15. (emergency medicine or ed or er or a&e).ti,ab.

16. 13 or 14 or 15

17. 12 or 16

18. *practice guideline/

19. (guideline? or guidance?).ti,ab.

20. 18 or 19

21. (adhere* or non-adhere* or nonadhere* or concord* or non-concord* or nonconcord* or discord* or comply or complian* or non-complian* or noncomplian* or align* or nonalign* or nonalign* or congruen* or incongruen* or consisten* or inconsisten* or contradict*).ti,ab.

22. ((does or "does not" or doesn?t or did or "did not" or didn?t or "not" or fail*) adj3 (follow* or met or meet or meeting or match or matching or "in line with")).ti,ab.

23. ((follow* or met or meet or meeting or match or matching or "in line with" or keep or kept or keeping or utili?ation or utile?e? or change?) adj5 (criteria or recommend* or guideline? or guidance)).ti,ab.

24. clinical competence/ or nursing competence/

25. 21 or 22 or 23 or 24

26. 20 and 25

27. diagnostic procedure/ or exp blood examination/ or exp cardiovascular system examination/ or exp digestive system examination/ or exp endocrine system examination/ or exp neurologic examination/ or exp respiratory tract examination/ or exp urogenital system examination/

28. (diagnos* or detect* or test* or screen* or manag*).ti.

29. (imaging or radiolog* or tomogra* or ct scan* or pet scan* or echocardiogra* or angiogra* or ultrasound* or ultrasonogra*).ti,ab.

30. ((medical or clinical or diagnos* or screening or routine or laboratory) adj5 (test* or investigation?)).ti,ab.

31. ((h?ematolog* or blood or urin* or saliva*) adj5 test*).ti,ab.

32. ((stress* or physical or function*) adj5 test*).ti,ab.

33. 27 or 28 or 29 or 30 or 31 or 32

34. 17 and 26 and 33

35. ((necessary or unnecessary or appropriate* or inappropriate* or waste* or utili?ation or indicated or excess* or less or more or increas* or decreas*) adj10 (test* or imaging or radiolog* or tomogra* or ct scan* or pet scan* or echocardiogra* or angiogra* or ultrasound* or ultrasonogra* or investigation?)).ti,ab.

36. ((order* or request*) adj5 (test* or imaging or radiolog* or tomogra* or ct scan* or pet scan* or echocardiogra* or angiogra* or ultrasound* or ultrasonogra* or investigation?)).ti,ab.

37. Unnecessary Procedures/

38. 35 or 36 or 37

39. 17 and 20 and 38

40. 17 and 33 and 37

41. 34 or 39 or 40

42. limit 41 to yr="1999 -Current"

43. limit 42 to english language

44. (exp animals/ or nonhuman/) not human/

45. 43 not 44

46. conference*.pt.

47. 45 and 46

48. 45 not 46

49. exp child/ not (exp Child/ and exp Adult/)

50. 48 not 49

51. 48 not 49

52. limit 47 to yr="2015 -Current"

6.3 Risk of bias

	Was the study"s target population a close representation of the national population in relation to relevant variables?	Does the inclusion criteria match the target population of guideline ?	Were all eligible participants included in the study?	Was the likelihood of non-response bias <2High% ?	Was an acceptable disease, test or symptom definition used?	Was data extracted/collected in an objective way?	Was the interval from symptoms to test clinically appropriate for the diagnosis of interest?	Did they report extractable measures ?	Other bias ?
Ahmad2012	Low	Unclear	Low	Unclear	Low	Unclear	Unclear	Low	Low
Aljebreen 2013	Low	Low	Low	Low	Unclear	Unclear	Low	Low	High
Azzam 2015	Low	Low	Unclear	High	Low	Unclear	Low	Low	Low
Belletti2013	Low	Unclear	Unclear	Unclear	Low	Low	Low	Low	Low
Bertella 2013	Low	Low	Low	Low	Unclear	Low	Unclear	Low	Low
Bhatt 2001	Low	Low	Low	High	High	Unclear	Unclear	Low	High
Bishop 2003	High	Low	Low	Low	Low	Low	Low	Low	Low
Cai 2015	Low	Low	Unclear	Low	Unclear	Unclear	Unclear	Low	Low
Caplan 2000	Low	Low	Low	Low	Unclear	Unclear	Unclear	Low	Low
Cardin 2005	Low	Low	Low	Low	Low	Low	Unclear	Low	Low
Cardin 2007	Low	Low	Low	Unclear	Low	Low	Low	Low	Low
Chan 2004	Low	Low	Low	Low	Low	Low	Unclear	Low	Low
Chan 2006	Low	Low	Low	High	Unclear	Low	Unclear	Low	Low
Chavez 2009	Low	Low	Low	Low	Low	Low	High	Low	High
Droogendijk 2011	Unclear	Low	High	Unclear	Low	Unclear	Low	Low	Low
Eccles 2001	Low	Low	Low	Low	Unclear	Unclear	Unclear	Low	Low
Elwyn 2007	Low	Low	Unclear	Unclear	Unclear	Low	Low	High	High
Gerrits2008	Low	Unclear	High	Low	Low	Low	Unclear	Low	Low
Gibbons 2010c	Low	Low	Low	Low	Low	High	Low	Low	Low
Girard 2010	High	Low	Unclear	High	Unclear	High	Unclear	Low	High
Gnani 2004	Low	Low	Unclear	Low	Unclear	Low	Unclear	Low	Low
Grover 2007	Low	Low	Unclear	Low	Low	High	Unclear	Low	Low
Gurzun 2014	Low	Low	High	Low	High	Low	Unclear	Low	High
Hassan 2007	Low	Low	Low	High	Unclear	Low	Unclear	Low	Low
Hughes-Anderson 2002	High	Low	Unclear	Low	Unclear	Unclear	unclear	Low	Low
Ip2014	Low	Low	High	Unclear	Low	Unclear	Unclear	Low	High
Johnson 2011	Low	Unclear	Unclear	Low	High	Low	Low	Low	Low
Kovacs 2013	Low	Unclear	High	Low	Low	Low	Unclear	Low	High
Lalude 2014	Low	Low	Low	Low	High	Low	Unclear	Low	High
Landry 2011	Low	Unclear	Low	Low	Unclear	Unclear	Unclear	Low	Low

Lange 2007	Low	Low	Unclear	Low	Unclear	Unclear	Unclear	Low	Low
Linder 2006	Low	High	Unclear	Low	Low	Low	Low	Unclear	Low
Lipczynska 2012	High	High	Unclear	Low	Low	Unclear	Low	Low	High
Llor 2011	Low	Low	Low	High	Low	Low	Low	Low	Low
Loo 2009	Low	Low	Low	Low	Low	Low	Unclear	Low	Low
Mafi2013	Low	Low	Unclear	Unclear	Low	Low	Unclear	Low	Low
Mafi2015	Low	Low	Unclear	Unclear	Low	Low	Unclear	Low	Low
Majumdar 2003	Low	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Michaleff 2012	Low	Low	Low	Unclear	High	Low	Unclear	Low	Low
Moscavitch 2009	Low	Low	Unclear	Low	Low	Unclear	Low	Low	Low
Musicco 2004	High	Low	Low	Unclear	Unclear	Unclear	Unclear	High	High
Nicholson 2010	Low	Low	Low	Low	Unclear	Low	Low	Low	Low
Nicopoullos 2003	High	High	Low	High	Low	Unclear	Unclear	Low	High
Noya 2008	Low	Low	Low	Low	Unclear	Unclear	Unclear	Low	High
Piccoliori 2013	Low	Low	Low	Low	Low	Unclear	Low	Low	High
Pimlott 2006	Low	Low	Unclear	High	Unclear	Unclear	Unclear	Low	Low
Piterman 2008	Low	Unclear	Low	Unclear	Unclear	Low	Unclear	High	High
Remedios 2014	Low	Unclear	Low	High	Unclear	Low	Unclear	Low	High
Schers 2000	Low	Low	Low	Unclear	Unclear	Low	Unclear	Low	Low
Smith 2008	Low	Low	Low	Low	Unclear	Low	Unclear	Low	Low
Sokol 2015	Low	Low	Low	Low	Low	Low	High	Low	High
Ulrik 2010	Low	Low	Low	High	Low	Low	unclear	Low	High
Ulrik 2013	Low	Low	Low	High	Low	Low	unclear	Low	Low
van Gurp 2013	Low	Low	Low	Low	Unclear	Low	Unclear	Low	Low
Williams2010	Low	Low	Unclear	Low	Low	Unclear	Unclear	Low	Low

6.4: Table 1

Study	Country	Length (days)	N (men%)	Population	Test
Under-use					
Ahmad 2012	Indonesia	181	554 (41%)	Patients registered at health clinics where TB was suspected	Sputum smear microscopy
Belletti 2013	USA	N/S	1517 (46%)	Patients with COPD	Pulmonary function tests (PFT)
Bertella 2013	Italy	1765	437 (286)	Patients with COPD	PFTs
Caplan 2000	USA	365	81	Patients who under went FNA of thyroid	FNA of thyroid
Chavez 2009	USA	2920	200 (48%)	Patients with COPD	PFT
Droogendijk 2011	Netherlands	730	287 (45%)	Women >50yrs and men >18 years with Iron Deficiency Anaemia	Upper endoscopy and colonoscopy
Gerrits 2008	Netherlands	2556	65 (0%)	Women aged 18 – 65 yrs with newly diagnosed urinary incontinence	Urine dipstick
Gibbons 2010	New Zealand	364	265	Patients with subclinical hypothyroidism	Free T4
Gnani 2004	UK	365	90 (53%)	Patients with heart failure	CXR, ECG and Echocardiogram
Girard 2010	France	28	19 (37%)	Patients with acute hepatitis	Hepatitis serology (HBs antigens, anti-HBc anitbodies)
Hughes-Anderson 2002a	Australia	1613	4400 (55%)	Patients who had colonoscopy	Colonoscopy

Lange 2007	Denmark	91	2549 (44%)	Patients with COPD	PFTs
Lipczynska 2012	Poland	61	93	Aged ≥ 55 with Heart Failure (HF) or HF risk factors	Echocardiogram, BNP, CXR
Loo 2009	UK	364	131 (50%)	Patients with Atrial Fibrillation	Echocardiogram
Majumdar 2003	USA	2371	1130 (47%)	Patients with dyspepsia and patients with peptic ulcer disease (PUD)	Upper endoscopy and H.pylori
Moscavitch 2009	Brazil	61	167 (43%)	Patients with Heart Failure	ECG, CXR, Echocardiogram
Musicco 2004	Italy	N/S	1549 (38%)	Patients being assessed for Dementia	Collection of laboratory tests to rule out conditions with similar presenting symptoms to dementia
Nicholson 2010	UK	1827	6943 (100%)	Men with epididymo-orchtitis	C. trachomatis, N. gonorrhoeae, urethral swabs and midstream urinalysis.
Nicopoullos 2003	UK	242	32	Patients with subfertility	Mid-luteal progesterone and semen analysis
Pimlott 2006	Canada	1611	160 (34%)	Patients with Dementia	FBC, TSH, serum electrolytes, serum calcium, glucose
Smith 2008	UK	731	29870 (52%)	Patients with COPD	PFT
Sokol 2015	USA	3652	75902 (23%)	Patients with Asthma	PFT
Ulrik 2010	Denmark	121	1716 (44%)	Patients with COPD	PFT
Ulrik 2013	Denmark	731	4058	Patients with COPD	PFT
Over-use					
Aljebreen 2013	Saudi	365	147 (51%)	Patients who had upper endoscopy	Upper endoscopy

	Arabia				
Azzam 2015	Saudi Arabia	121	161 (30%)	Dyspeptic patients who had upper endoscopy	Upper endoscopy
Bhatt 2001	UK	504	437 (65%)	Patients referred for pelvis x-rays	Pelvis x-ray
Bishop 2003	Canada	28	139	Patients with non-red flag LBP	Advanced imaging (CT, MRI or bone scan)
Cai 2015	USA	121	550 (46%)	Patients who under went upper endoscopy	Upper endoscopy
Chan 2004	Malaysia	153	250 (45%)	Patients who under went upper endoscopy	Upper endoscopy
Chan 2006	Malaysia	184	27 (63%)	Patients who underwent 'diagnostic colonoscopies'	Colonoscopy
Cardin 2005	Italy	151	1678	Patients with uninvestigated dyspepsia	Upper endoscopy
Cardin 2007	Italy	182	1/04/2001	Dyspeptic patients who had upper endoscopy	Upper endoscopy
Eccles 2001	UK	182	275	Patients who had knee or lumbar x-ray	Lumbar or knee x-ray
Elwyn 2007	UK	184	215	Patients who under went upper endoscopy	Upper endoscopy
Grover 2007	USA	364	68 (0%)	Patients with uncomplicated UTI	Urine culture and sensitivity analysis
Gurzun 2014	UK	7	1070 (54%)	Echocardiogram	Echocardiogram
Hassan 2007	Italy	30	3769 (46%)	Patients who under went upper endoscopy	Upper endoscopy
Hughes-Anderson 2002b	Australia	1613	4400 (55%)	Patients who had upper endoscopy,	Upper endoscopy
Ip 2014	USA	1096	100 (43%)	Patients with non-red flag LBP	MRI lumbar spine
Johnson 2011	USA	510	779 (0%)	Patients with uncomplicated UTI	Urine culture
Kovacs 2013	Spain	183	602 (48%)	Patients with non-red flag LBP	MRI lumbar spine

Lalude 2014	USA	121	102	Patients who had SPECT Myocardial perfusion imaging (MPI) studies	Single Photon Emission CT (SPECT) MPI
Landry 2011	USA	272	124	Patients who had U/S of thyroid, pelvis, abdo, carotid or soft tissue	Thyroid, pelvis, abdomen, carotid or soft tissue ultrasound
Linder 2006	USA	608	1076 (19%)	Patients with pharyngitis	Strep testing (rapid antigen detection test, throat culture)
Llor 2011	Spain	122	658 (0%)	Women with UTI	Urine cultures
Mafi 2013	USA	4377	8066	Patients with non-red flag LBP	X-ray, CT or MRI
Mafi 2015	USA	4018	9362 (25%)	Patients with uncomplicated headache (non-red flag	CT and MRI
Michaleff 2012	Australia	3621	3070 (70%)	Patients reporting first time neck pain or LBP (non-specific, non red flag)	Any radiological test
Noya 2008	Israel	N/S	209 (35%)	Patients who had H.pylori testing	H. pylori test
Piccoliori 2013	Italy	63	475 (43%)	Acute or chronic non-red flag LBP	Any radiological test
Piterman 2008	Australia	550	19219	Patients with GORD	Endoscopy. Barium Swallow
Remedios 2014	UK	N/S	2026	Patients who had CTs and/or MRIs	CT and/or MRI
Schers 2000	Netherlands	214	1096 (50%)	Patients with non-red flag LBP	X-ray
Van Gurp 2013	Netherlands	366	155 (38%)	Patients who had Echocardiogram	Echocardiogram
Williams 2010	Australia	1005	1706 (43%)	Patients with non-red flag LBP	All imaging

6.5: Table 2

Study	Test	Guideline authority and recommendation	Measure of inappropriateness (95%CI)
Under-use			
Girard 2010	Hepatitis B serology	Ministry of Health (France): Hepatitis serology for suspected acute hepatitis	100% (82.4 to 100%)
Nicholson 2010	Neisseria Gonorrhoea serology	CDC (US)/British Association for Sexual Health and HIV: Test for N. gonorrhoea for suspected Epididymitis	99.9% (99.85 to 99.98%)
Nicholson 2010	Chlamydia Trachomatis	CDC (US)/British Association for Sexual Health and HIV: Test for C. Trachomatis for suspected Epididymitis	97.4% (97.0 to 97.8%)
Nicopoullos 2003	Semen Analysis	Royal College of Obstetricians and Gynaecologists: Semen analysis for Infertility	96.9% (95%CI: 83.8 to 99.9%)
Lipczynska 2012	Brain Natriuretic Peptide (BNP)	European Society Cardiology: BNP for Heart Failure	95.7% (95%CI: 89.4 to 98.8%)
Musicco 2004	Collection of laboratory tests	European Federation of Neurological Societies: Collection of laboratory tests to rule out conditions with similar presenting symptoms to dementia	93.42% (95%CI: 92.1 to 94.6%)
Nicholson 2010	Urethral swabs	CDC (US)/British Association for Sexual Health and HIV: Urethral swabs for suspected epididymitis (Urethral swabs)	90.7% (95%CI:89.9 to 91.3%)
Moscavitch 2009	Echocardiogram	Brazilian Society of Cardiology: Echocardiography for Heart Failure	88.6% (95%CI: 82.8 to 93.0%)
Majumdar 2003	Upper Endoscopy	American Gastroenterological Association: Appropriate use of	81.2% (78.8 to 83.4%)

		Upper Endoscopy for Dyspepsia	
Nicholson 2010	Mid stream	CDC (US)/British Association for Sexual Health and HIV: Midstream urinalysis for suspected Epididymitis	78.2 (95%CI: 77.3 to 79.3%)
Ulrik 2013	Pulmonary function tests (PFTs)	Danish National Board of Health: PFTs to diagnosis COPD	78.1 (76.8% to 79.4%)
Sokol 2015	Pulmonary function tests (PFTs)	National Asthma Education and Prevention Program (US): PFTs for asthma	76.5% (64.6 to 85.9%)
Belletti 2013	Pulmonary function tests (PFTs)	Global Initiative for Chronic Obstructive Lung Disease (GOLD): PFTs for COPD	73.0% (10.7 to 75.3%)
Ahmad 2012	Tuberculosis smear	World Health Organisation: Smear for suspected TB	72.4% (68.5 to 76.1%)
Droogendijk 2011	Colonoscopy	Ministry of Health (Netherlands): Colonoscopy for unexplained Iron Deficiency Anaemia	68.6% (62.9 to 74.0%)
Pimlott 2006	Serum Calcium	Canadian Consensus Conference on Dementia: Serum Calcium for Dementia	65.0 (57.1 to 72.4%)
Gerrits 2008	Urine dip stick	NICE: Urine dip stick for urinary incontinence	60.0% (47.1 to 72.0%)
Nicopoullos 2003	Mid-luteal progesterone	Royal College of Obstetricians and Gynaecologists: Mid-luteal progesterone for Infertility	59.4% (40.6 to 76.3%)
Gnani 2004	Echocardiogram	Department of Health (UK): Echocardiogram for Heart Failure	57.8% (46.1 to 68.1%)
Loo 2009	Echocardiogram	ACC, AHA, ESC: Echocardiogram to identify causes or complications of atrial fibrillation	55.7% (46.8 to 64.39%)
Ulrik 2010	Pulmonary function tests (PFTs)	Global Initiative for Chronic Obstructive Lung Disease (GOLD): PFTs for COPD	55.0% (52.6 to 57.4%)

Lipczynska 2012	Echocardiogram	European Society Cardiology: Echocardiogram for Heart Failure	53.8% (43.1 to 64.2%)
Gnani 2004	ECG	Department of Health (UK): ECG for Heart Failure	51.1% (40.4% to 61.8%)
Moscavitch 2009	ECG	Brazilian Society of Cardiology: ECG for Heart Failure	46.1 (38.4 to 54.0)
Moscavitch 2009	Chest X-ray	Brazilian Society of Cardiology: CXR for Heart Failure	44.9% (37.2 to 52.8%)
Gibbons 2010	Thyroid function tests	New Zealand Best Practice: Appropriate use of Thyroid Function tests	44.9% (38.8 to 51.1%)
Chavez 2009	Pulmonary function tests (PFTs)	Global Initiative for Chronic Obstructive Lung Disease (GOLD): PFTs for COPD	41.5% (34.6 to 48.7%)
Lange 2007	Pulmonary function tests (PFTs)	Global Initiative for Chronic Obstructive Lung Disease (GOLD): PFTs for COPD	40.0% (38.1 to 42.0)
Smith 2008	Pulmonary Function Tests (PFTs)	NICE: PFTs for COPD	38.1% (37.5 to 38.6%)
Pimlott 2006	Glucose testing	Canadian Consensus Conference on Dementia: Glucose testing for Dementia	36.9% (29.4% to 44.9%)
Gnani 2004	Chest X-ray	Department of Health (UK): CXR for Heart Failure	36.7% (26.8 to 47.5%)
Majumdar 2003	H.pylori	American Gastroenterological Association/American College of Gastroenterology: appropriateness of H.pylori test	34.4% (28.9 to 40.3%)
Pimlott 2006	Thyroid Stimulating Hormone (TSH)	Canadian Consensus Conference on Dementia: TSH for dementia	33.1% (25.9 to 41.0%)
Lipczynska 2012	Chest x-ray (CXR)	European Society Cardiology: CXR for Heart Failure	25.8% (17.3 to 35.9%)
Bertella 2013	Pulmonary function tests (PFTs)	Global Initiative for Chronic Obstructive Lung Disease (GOLD): PFTs for COPD	25.6% (21.6 to 30.0%)

Pimlott 2006	Serum electrolytes	Canadian Consensus Conference on Dementia: Serum electrolytes for dementia	21.9% (15.7 to 29.1%)
Pimlott 2006	Full Blood Count (FBC)	Canadian Consensus Conference on Dementia: FBC for dementia	18.1% (12.5 to 25.0%)
Caplan 2000	Fine needle aspiration (FNA) of thyroid	American Thyroid Association/American Association of Clinical Endocrinologists: FNA for thyroid nodules	9.9% (4.4 to 18.5%)
Hughes-Anderson 2002a	Colonoscopy	American Society for Gastrointestinal Endoscopy: Appropriateness of Colonoscopy	8.2% (5.3 to 12.1%)
Over-use			
Piterman 2008	Barium Swallow	Gastroenterological Society of Australia: Barium Swallow for GORD	94.20% (95%CI: 93.9 to 94.5%)
Gurzun 2014	Echocardiogram	American College of Cardiology: Appropriate use of Echocardiography	92.0% (95%CI: 90.2% to 93.5%)
van Gurp 2013	Echocardiogram	Netherlands Society of Cardiology: Appropriate use of Echocardiography	76.7% (76.4 to 77.0%)
Grover 2007	Urine cultures	Infectious Disease Society of America: Urine cultures not required for uncomplicated UTI diagnosis	76.5% (64.6 to 85.9%)
Eccles 2001	Knee x-ray	Royal College of Radiologists: No x-ray for knee pain without restriction of movement	74.7% (69.6 to 79.3%)
Cardin 2005	H. Pylori breath test	European Society of Primary Care Gastroenterology:	74.4% (58.8 to 86.5%)

		Appropriate use of H. pylori	
Johnston 2011	Urine cultures	European Association of Urology: Urinary cultures not required for uncomplicated urinary tract infections	57.4% (53.8 to 60.9%)
Bhatt 2001	Hip x-ray	Royal College of Radiologists (UK): No hip x-ray for hip pain without restriction of movement	57.2% (52.5 to 61.8%)
Eccles 2001	Lumbar spine x-ray	Royal College of Radiologists (UK): no x-ray for non-red flag LBP	56.4% (50.3 to 62.3%)
Piterman 2008	Upper endoscopy	Gastroenterological Society of Australia: Upper endoscopy for GORD	53.5% (52.8 to 54.2%)
Chan 2006	Colonoscopy	American Society for Gastrointestinal Endoscopy: Appropriateness of Colonoscopy	51.9% (32.0 to 71.3%)
Aljebreen 2013	Upper endoscopy	The American Society for Gastrointestinal Endoscopy: Appropriateness of Upper Endoscopy	46.9% (38.7 to 55.3%)
Elwyn 2007	Upper endoscopy	NICE: Appropriate tests for dyspepsia	45.1% (38.3 to 52.0%)
Noya 2008	H.Pylori	The European Helicobacter Study Group: Appropriate use of H. pylori	44.5 (37.6 to 51.5%)
Cardin 2005	Upper endoscopy	European Society of Primary Care Gastroenterology: Upper Endoscopy for H.pylori	44.1% (35.9 to 52.6%)
Cardin 2007	Upper endoscopy	European Society of Primary Care Gastroenterology: Upper Endoscopy for H.pylori	41.9% (38.3 to 45.5%)
Cai 2015	Upper endoscopy	American College of Physicians: Upper endoscopy for GORD	37.7% (33.8 to 42.0%)

Azzam 2015	Upper endoscopy	American Gastroenterological Association: Upper Endoscopy for Dyspepsia	36.7 (29.2 to 44.6%)
Llor 2011	Urine cultures	European Association of Urology: Urinary cultures not required for uncomplicated urinary tract infections	35.9% (32.2 to 40.0%)
Hassan 2007	Upper endoscopy	The American Society for Gastrointestinal Endoscopy: Appropriateness of Upper Endoscopy	29.4% (28.0 to 30.9%)
Landry 2011	Carotid ultrasound	Canadian Association of Radiologists 2005 guidelines: Carotid U/S	25.0% (17.7 to 33.6%)
Piccoliori 2013	Lumbar spine radiology (all)	Ministry of Health (Italy): No imaging for non-red flag LBP	24.0% (20.2 to 28.1%)
Michaleff 2012	Lumbar spine x-ray	National Health and Medical Research Council (Australia) (NHMRC): No x-ray for non-red flag LBP	24.0% (22.9 to 25.1%)
Williams 2010	Lumbar spine radiology (all)	National Health and Medical Research Council (Australia) (NHMRC): No imaging for non-red flag LBP	23.9% (21.9 to 26.0%)
Michaleff 2012	Cervical spine x-ray	Australian National Health and Medical Research Council: No x-ray for neck pain	22.8% (21.3 to 24.3%)
Ip 2014	Lumbar spine MRI	American College of Physicians/American Pain Society: no MRI for non-red flag LBP	22.0% (14.3 to 31.4%)
Williams 2010	Lumbar spine x-ray	National Health and Medical Research Council (Australia) (NHMRC): No x-ray for non-red flag LBP	20.2% (18.3 to 22.2%)
Landry 2011	Thyroid ultrasound	Canadian Association of Radiologists 2005 guidelines: Thyroid U/S	19.0% (12.1 to 27.0%)

Lalude 2014	Single Photon Emission Computed Tomography	American College of Cardiology: SPECT for chest pain	18.6% (11.6 to 27.6%)
Linder 2006	Streptococcal throat cultures	American College of Physicians/Infectious Disease Society of America: Pharyngitis	15.0 (12.9 to 17.2%)
'Mafi 2013	Lumbar spine x-ray	American College of Physicians/American Pain Society: no x-ray for non-red flag LBP: *2009-2010*	13.0% (11.1 to 15.1%)
		2007-2008	12.9% (11.1 to 14.9%)
		2005-2006	12.8% (11.0 to 14.8%)
		2003-2004	12.3% (10.7 to 14.0%)
		2001-2002	12.0% (10.3 to 13.8%)
		1999 - 2000	11.8% (10.2 to 13.6%)
Landry 2011	Abdominal ultrasound	Canadian Association of Radiologists 2005 guidelines: Abdominal U/S	12.1% (6.9 to 19.2%)
Mafi 2015	CT or MRI Brain	The American Headache Society/American Academy of Neurology for Choosing Wisely: No CT or MRI for non-red flag headache *2009 - 2010*	13.9% (12.2 to 15.7%)
		2007 – 2008	13.5% (11.8 to 15.3%)
		2005 – 2006	9.4% (8.0 to 11.0%)
		2003 – 2004	7.5% (6.3 to 8.9%)
		2001 – 2002	7.1% (5.9 to 8.4%)
		1999 - 2000	6.7% (5.4 to 8.2%)
Kovacs 2013	Lumbar spine radiology tests (all)	NICE, ACP: No imaging for LBP	12.0% (9.5 to 14.8%)

Chan 2004	Upper endoscopy	The American Society for Gastrointestinal Endoscopy: Appropriateness of Upper Endoscopy	10.4% (6.9 to 14.9%)
Hughes-Anderson 2002b	Upper endoscopy	The American Society for Gastrointestinal Endoscopy: Appropriateness of Upper Endoscopy	7.5% (4.7 to 11.1%)
Remedios 2014	CT (any)	Royal College of Radiologists (UK): CT	6.9% (5.8 to 8.1%)
	MRI (any)	Royal College of Radiologists (UK): MRI	5.2% (4.1 to 6.5%)
Bishop 2003	Lumbar spine radiology tests (all)	Workers Compensation Board of British Columbia: No imaging for non-red flag LBP	5.0% (2.1 to 10.1%)
Williams 2010	Lumbar spine CT	National Health and Medical Research Council (Australia) (NHMRC): No CT for non-red flag LBP	3.7% (2.9 to 4.7%)
Schers 2000	Lumbar spine radiology tests (all)	The Netherlands College of General Practitioners: No imaging for non-red flag LBP	3.1% (2.2 to 4.3%)
Landry 2011	Soft tissue ultrasound	Canadian Association of Radiologists 2005 guidelines: Soft tissue U/S	2.4% (0.5 to 6.9%
Landry 2011	Pelvic ultrasound	Canadian Association of Radiologists 2005 guidelines: Pelvic U/S	1.6% (0.2 to 5.7%)
Williams 2010	Lumbar spine Ultrasound	National Health and Medical Research Council (Australia) (NHMRC): No U/S for non-red flag LBP	0.59% (0.28 to 1.1%)
Williams 2010	Lumbar spine MRI	National Health and Medical Research Council (Australia) (NHMRC): No MRI for non-red flag LBP	0.18% (0.04 to 0.5%)

Appendix Chapter 7

7.1 AGREE II Scores

Guideline	Guideline recommendation	Evidence supporting recommendation	Prevalence inappropriate (%)	AGREE II score	Domain 1 AGREE II (%)	Domain 2 AGREE II (%)	Domain 3 AGREE II (%)	Domain 4 AGREE II (%)	Domain 5 AGREE II (%)	Domain 6 AGREE II (%)
Australian NHMRC: Evidence-Based Management of Acute Musculoskeletal Pain[1,2]	No MRI for non-red flag low back pain	Sorby W 1989[a]	0.18	6	100.0	100.0	99.0	91.7	75.0	45.8
Australian NHMRC: Evidence-Based Management of Acute Musculoskeletal Pain[1,2]	No CT for non-red flag low back pain	Wiesel 1986[b]	3.7	6	100.0	100.0	99.0	91.7	75.0	45.8
Australian NHMRC: Evidence-Based Management of Acute Musculoskeletal Pain[1,2]	No x-rays for non-red flag low back pain	Kerry 2002[c] Kendrick 2001[d]	20.2	6	100.0	100.0	99.0	91.7	75.0	45.8
Australian NHMRC: Evidence-Based Management of Acute Musculoskeletal Pain[1,2]	No x-rays for non-red flag low back pain	Kerry 2002[c] Kendrick 2001[d]	24.0	6	100.0	100.0	99.0	91.7	75.0	45.8

Australian NHMRC: Evidence-Based Management of Acute Musculoskeletal Pain[1,2]	No imaging for non-red flag neck pain	Heller 1983[e] Johnson 1997[f]	22.8	6	100.0	100.0	99.0	91.7	75.0	45.8
Australian NHMRC: Evidence-Based Management of Acute Musculoskeletal Pain[1,2]	No imaging for non-red flag low back pain	Kerry 2002[c] Kendrick 2001[d]	23.9	6	100.0	100.0	99.0	91.7	75.0	45.8
American College of Physicians and American Pain Society: Diagnosis and Treatment of Low Back Pain[3]	No x-rays for non-red flag low back pain	Kerry 2002[c] Kendrick 2001[d]	11.8	5	97.2	55.6	85.4	97.2	41.7	66.7
American College of Physicians and American Pain Society: Diagnosis and Treatment of Low Back Pain[3]	No x-rays for non-red flag low back pain	Kerry 2002[c] Kendrick 2001[d]	12.0	5	97.2	55.6	85.4	97.2	41.7	66.7
American College of Physicians and American Pain Society: Diagnosis and Treatment of Low Back Pain[3]	No x-rays for non-red flag low back pain	Kerry 2002[c] Kendrick 2001[d]	12.3	5	97.2	55.6	85.4	97.2	41.7	66.7

American College of Physicians and American Pain Society: Diagnosis and Treatment of Low Back Pain[3]	No x-rays for non-red flag low back pain	Kerry 2002[c] Kendrick 2001[d]	12.8	5	97.2	55.6	85.4	97.2	41.7	66.7
American College of Physicians and American Pain Society: Diagnosis and Treatment of Low Back Pain[3]	No x-rays for non-red flag low back pain	Kerry 2002[c] Kendrick 2001[d]	12.9	5	97.2	55.6	85.4	97.2	41.7	66.7
American College of Physicians and American Pain Society: Diagnosis and Treatment of Low Back Pain[3]	No x-rays for non-red flag low back pain	Kerry 2002[c] Kendrick 2001[d]	13.0	5	97.2	55.6	85.4	97.2	41.7	66.7
American College of Physicians and American Pain Society: Diagnosis and Treatment of Low Back Pain[3]	No MRI for non-red flag low back pain	Gilbert 2005[g]	22.0	5	97.2	55.6	85.4	97.2	41.7	66.7
The American Headache Society/American Academy of Neurology for Choosing Wisely[4,5]	No CT or MRI for non-red flag headache	Expert Opinion	6.7	4	58.3	38.9	53.1	77.8	33.3	70.8

The American Headache Society/American Academy of Neurology for Choosing Wisely[4,5]	No CT or MRI for non-red flag headache	Expert Opinion	7.1	4	58.3	38.9	53.1	77.8	33.3	70.8
The American Headache Society/American Academy of Neurology for Choosing Wisely[4,5]	No CT or MRI for non-red flag headache	Expert Opinion	7.5	4	58.3	38.9	53.1	77.8	33.3	70.8
The American Headache Society/American Academy of Neurology for Choosing Wisely[4,5]	No CT or MRI for non-red flag headache	Expert Opinion	9.4	4	58.3	38.9	53.1	77.8	33.3	70.8
The American Headache Society/American Academy of Neurology for Choosing Wisely[4,5]	No CT or MRI for non-red flag headache	Expert Opinion	13.5	4	58.3	38.9	53.1	77.8	33.3	70.8
The American Headache Society/American Academy of Neurology for Choosing Wisely[4,5]	No CT or MRI for non-red flag headache	Expert Opinion	13.9	4	58.3	38.9	53.1	77.8	33.3	70.8
AmericanThyroid Association: Guidelines for patients with Thyroid Nodules[6]	TSH blood test and FNA cytology for the initial evaluation of thyroid nodules	Expert Opinion	9.9	1.5	63.9	19.4	10.4	36.1	0.0	0.0

Infectious Disease Society of America: Practice Guidelines for the Diagnosis and Management of Group A Streptococcal Pharyngitis[7]	Don't do strept testing if Centor Criteria is 0, 1 or 2	Centor 1981[h] Dagnelie 1998[i]	91.5	4	97.2	19.4	42.7	75.0	20.8	0.0
Canadian Consensus Conference on Dementia[8]	Complete Blood Count (CBC) is required for patients with a new diagnosis of Dementia	Expert Opinion	18.1	5	91.7	63.9	60.4	77.8	20.8	62.5
Canadian Consensus Conference on Dementia[8]	Serum Electrolytes is required for patients with a new diagnosis of Dementia	Expert Opinion	21.9	5	91.7	63.9	60.4	77.8	20.8	62.5
Canadian Consensus Conference on Dementia[8]	Thyroid Stimulating Hormone (TSH) is required for patients with a new diagnosis of Dementia	Expert Opinion	33.1	5	91.7	63.9	60.4	77.8	20.8	62.5
Canadian Consensus Conference on Dementia[8]	Blood glucose is required for patients with a new diagnosis of Dementia	Expert Opinion	36.9	5	91.7	63.9	60.4	77.8	20.8	62.5
Canadian Consensus Conference on Dementia[8]	Serum Calcium is required for patients with	Expert Opinion	65.0	5	91.7	63.9	60.4	77.8	20.8	62.5

	a new diagnosis of Dementia									
American College of Cardiology: Appropriate Use Criteria for Cardiac Radionuclide Imaging[9]	Numerous: Do not use Cardiac Radionuclide for: Low pretest probability of CAD, ECG interpretable AND able to exercise, Intermediate CHD risk (ATP III risk criteria), Definite ACS, No clinical risk factors, Low CHD risk (ATP III risk criteria), Known CAD on coronary angiography or prior abnormal stress imaging study, Last stress imaging study done less than 2 years ago, Less than 2 years after PCI	Expert Opinion	18.6	4	58.3	30.6	57.3	77.8	27.1	62.5
European Society of Cardiology Guidelines for diagnosis and treatment of acute and chronic heart	Chest x-ray to confirm or refute the diagnosis of heart failure	Expert Opinion	25.8	4.5	91.7	38.9	60.4	75.0	45.8	62.5

failure[10]										
European Society of Cardiology Guidelines for diagnosis and treatment of acute and chronic heart failure[10]	Echocardiogram to confirm or refute the diagnosis of heart failure	Expert Opinion	53.8	4.5	91.7	38.9	60.4	75.0	45.8	62.5
European Society of Cardiology Guidelines for diagnosis and treatment of acute and chronic heart failure[10]	Brain Natriuretic Peptide (BNP) to confirm or refute the diagnosis of heart failure	Mueller 2004[J]	95.7	4.5	91.7	38.9	60.4	75.0	45.8	62.5
American College of Gastroenterology: Medical Treatment of Peptic Ulcer Disease[11]	H.Pylori test to investigate dyspepsia	Cutler 1995[k]	34.4	3.5	69.4	22.2	43.8	61.1	18.8	8.3

American College of Physicians: Upper endoscopy for Gastroesophageal reflux disease[12]	Upper endo for any of: (1) nondysplastic Barrett esophagus without an EGD for surveillance in the past 3 years, (2) acute symptoms (<5 years) that persisted despite 4 to 8 weeks of twice-daily proton-pump-inhibitor therapy, (3) chronic symptoms (>5 years) in men who were older than 50 years, (4) alarm symptoms (dysphagia, bleeding, anemia, weight loss, and recurrent vomiting), and (5) severe erosive esophagitis despite 2 months of proton-pump inhibitor therapy	Expert Opinion	37.8	3.5	80.6	38.9	25.0	86.1	37.5	83.3
NICE: National clinical guideline on management of chronic obstructive pulmonary disease in adults in primary and	Pulmonary Function Tests to confirm or refute the diagnosis of COPD	Expert Opinion	38.1	7	100.0	100.0	94.8	97.2	72.9	91.7

secondary care[13]										
Best Practice New Zealand: Investigating Thyroid Function	Patients with subclinical hypothyroidism order a Free Thyroxine test	Expert Opinion	44.9	1.5	80.6	16.7	19.8	86.1	10.4	0.0
NICE: Managing dyspepsia in adults in primary care	In patients with dyspepsia, test and treat for H. pylori before endoscopy	Bytzer 1994[L] Lewin van den Broek 2001[m]	45.1	6.5	97.2	94.4	92.7	100.0	79.2	95.8
European Association of Urology: Guidelines on urinary and male Genital tract infections	No urine cultures for uncomplicated urinary tract infection	Expert Opinion	35.9	1.5	72.2	5.6	12.5	55.6	2.1	0.0
Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2001 Guideline[14]	Pulmonary Function Tests to confirm or refute the diagnosis of COPD	Expert Opinion	40.0	3.5	97.2	41.7	43.8	77.8	43.8	20.8
Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2001 Guideline[14]	Pulmonary Function Tests to confirm or refute the diagnosis of COPD	Expert Opinion	41.5	3.5	97.2	41.7	43.8	77.8	43.8	20.8
Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2007	Pulmonary Function Tests to confirm or refute the diagnosis of	Jackson 2003[n]	25.6	5	0.7	0.4	0.7	0.6	0.4	0.1

Guideline[15]	COPD									
Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2007 Guideline[15]	Pulmonary Function Tests to confirm or refute the diagnosis of COPD	Jackson 2003[n]	55.0	5	0.7	0.4	0.7	0.6	0.4	0.1
Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2007 Guideline[15]	Pulmonary Function Tests to confirm or refute the diagnosis of COPD	Jackson 2003[n]	73.0	5	0.7	0.4	0.7	0.6	0.4	0.1
American Gastroenterological Association: Evaluation of Dyspepsia[16]	For patients >50 years of age, order endoscopy for dyspepsia	Talley 1998[o]	81.2	3.5	77.8	41.7	44.8	75.0	8.3	8.3
U.S. National Institutes of Health: Guidelines for the Diagnosis and Management of Asthma[17]	Pulmonary Function Tests to confirm or refute the diagnosis of Asthma	Li 1996[p]	76.7	5.5	91.7	63.9	81.3	80.6	47.9	75.0
European Federation of Neurological Societies: Tests for Dementia[18]	CBC, electrolytes, calcium, glucose, renal and liver function tests and TSH are required for patients with a new diagnosis of Dementia	Expert Opinion	93.4	6	88.9	66.7	72.9	80.6	22.9	70.8

Canadian Task Force on the Periodic Health Examination[19]	Do not order TSH tests for the purpose of screening a patient who does not present with any symptoms or signs of primary hypothyroidism	Expert Opinion	22.4	5.5	100.0	27.8	74.0	97.2	14.6	91.7
American Academy of Otolaryngology: Clinical practice guideline on adult sinusitis[20]	Do not order CT scans for patients with uncomplicated Acute Sinusitis	Lau 1999[q]	0.6	6	100.0	72.2	84.4	97.2	66.7	83.3
American College of Gastroenterology. Guidelines for the management of dyspepsia[21]	4-8 weeks of PPI and H.Pylori testing and treatment before endoscopy in patients <55 years without alarm features	Expert Opinion	42.4	4.5	94.4	44.4	57.3	91.7	16.7	0.0
American Society For Gastrointestinal Endoscopy: Appropriate Use of Gastrointestinal Endoscopy[22]	20 indications for upper endoscopy (See Appendix 7.3)	Expert Opinion	10.4	2	86.1	66.7	24.0	66.7	4.2	0.0
American Society For Gastrointestinal Endoscopy: Appropriate	20 indications for upper endoscopy (See Appendix 7.3)	Expert Opinion	29.4	2	86.1	66.7	24.0	66.7	4.2	0.0

Use of Gastrointestinal Endoscopy[22]										
American Society For Gastrointestinal Endoscopy: Appropriate Use of Gastrointestinal Endoscopy[22]	20 indications for upper endoscopy (See Appendix 7.3)	Expert Opinion	46.9	2	86.1	66.7	24.0	66.7	4.2	0.0
American Society For Gastrointestinal Endoscopy: Appropriate Use of Gastrointestinal Endoscopy[22]	20 indications for upper endoscopy (See Appendix 7.3)	Expert Opinion	7.5	2	86.1	66.7	24.0	66.7	4.2	0.0
American Society For Gastrointestinal Endoscopy: Appropriate Use of Gastrointestinal Endoscopy[22]	20 indications for Colonoscopy (See Appendix 7.3)	Expert Opinion	8.2	2	86.1	66.7	24.0	66.7	4.2	0.0
American Society For Gastrointestinal Endoscopy: Appropriate Use of Gastrointestinal Endoscopy[22]	20 indications for Colonoscopy (See Appendix 7.3)	Expert Opinion	51.9	2	86.1	66.7	24.0	66.7	4.2	0.0

The European Society of Primary Care Gastroenterology: Management of H. pylori infections[23]	Upper endoscopy in patients with dysphagia and alarm symptoms (weight loss, vomiting, haematemesis, anaemia)	Expert Opinion	44.1	5.5	100.0	44.4	67.7	97.2	18.8	0.0
The European Society of Primary Care Gastroenterology: Management of H. pylori infections[23]	H. pylori testing should be considered for those patients who consult with recurrent symptoms of dyspepsia	Expert Opinion	74.4	5.5	100.0	44.4	67.7	97.2	18.8	0.0
CDC (US)/British Association for Sexual Health and HIV: Management of Epididymitis[24]	Midstream urinanalysis to rule out other causes	Expert Opinion	78.2	2.0	80.6	11.1	12.5	77.8	12.5	50.0
CDC (US)/British Association for Sexual Health and HIV: Management of Epididymitis[24]	Urethral swabs to rule out other causes	Expert Opinion	90.7	2.0	80.6	11.1	12.5	77.8	12.5	50.0
CDC (US)/British Association for Sexual Health and HIV: Management of	C. Trachomatis serology to rule out other causes	Expert Opinion	97.4	2.0	80.6	11.1	12.5	77.8	12.5	50.0

Epididymitis[24]										
CDC (US)/British Association for Sexual Health and HIV: Management of Epididymitis[24]	N. gonorrhoea serology to rule out other causes	Expert Opinion	99.9	2.0	80.6	11.1	12.5	77.8	12.5	50.0
Royal College of Obstetricians and Gynaecologists: Infertility[25]	Mid-luteal progesterone test for Infertility work-up	Wathen 1984[r]	59.4	2.0	80.6	11.1	12.5	77.8	12.5	50.0
Royal College of Obstetricians and Gynaecologists: Infertility[25]	Semen analysis for Infertility work-up	Opsahl 1996[s]	96.9	2.0	80.6	11.1	12.5	77.8	12.5	50.0
NICE: Urinary Incontinence[26]	Urine dipstick for Urinary Incontinence	Expert Opinion	60.0	6.0	100.0	61.1	84.4	97.2	81.3	37.5

7.2 Specific guideline recommendations (that were too long for Appendix Table 7.1)

Guideline: American Society For Gastrointestinal Endoscopy: Appropriate Use of Gastrointestinal Endoscopy[36]

Recommendation: Upper endoscopy for the following indications:

1. Upper abdominal symptoms, which persist despite an appropriate trial of therapy
2. Upper abdominal symptoms associated with other symptoms or signs suggesting serious organic disease (e.g., anorexia and weight loss) or in patients over 45 years of age.
3. Dysphagia or odynophagia.
4. Esophageal reflux symptoms, which are persistent or recurrent despite appropriate therapy.
5. Persistent vomiting of unknown cause.
6. Other diseases in which the presence of upper GI pathology might modify other planned management. Examples include, patients who have a history of ulcer or GI bleeding who are scheduled for organ transplantation, long-term anticoagulation or chronic nonsteroidal anti-inflammatory drug therapy for arthritis and those with cancer of the head and neck.
7. Familial adenomatous polyposis syndromes.
8. For confirmation and specific histologic diagnosis of radiologically demonstrated lesions: 1. Suspected neoplastic lesion. 2. Gastric or esophageal ulcer. 3. Upper tract stricture or obstruction.
9. Gastrointestinal bleeding: 1. In patients with active or recent bleeding. 2. For presumed chronic blood loss and for iron deficiency anemia when the clinical situation suggests an upper GI source or when colonoscopy is negative.
10. When sampling of tissue or fluid is indicated.
11. In patients with suspected portal hypertension to document or treat esophageal varices.
12. To assess acute injury after caustic ingestion.
13. Treatment of bleeding lesions such as ulcers, tumors,vascular abnormalities (e.g., electrocoagulation, heater probe, laser photocoagulation or injection therapy).
14. Banding or sclerotherapy of varices.
15. Removal of foreign bodies.
16. Removal of selected polypoid lesions.
17. Placement of feeding or drainage tubes (peroral, percutaneous endoscopic gastrostomy, percutaneous endoscopic jejunostomy).
18. Dilation of stenotic lesions (e.g., with transendoscopic balloon dilators or dilation systems employing guidewires).
19. Management of achalasia (e.g., botulinum toxin, balloon dilation).
20. Palliative treatment of stenosing neoplasms (e.g., laser, multipolar electrocoagulation, stent placement).

Recommendation: Colonoscopy for the following indications:

1. Evaluation of an abnormality on barium enema or other imaging study, which is likely to be clinically significant, such as a filling defect or stricture.
2. Evaluation of unexplained gastrointestinal bleeding. 1. Hematochezia 2. Melena after an upper GI source has been excluded. 3. Presence of fecal occult blood

3. Unexplained iron deficiency anemia.
4. Screening and surveillance for colonic neoplasia: 1. Screening of asymptomatic, average risk patients for colonic neoplasia. 2. Examination to evaluate the entire colon for synchronous cancer or neoplastic polyps in a patient with treatable cancer or neoplastic polyp. 3. Colonoscopy to remove synchronous neoplastic lesions at or around time of curative resection of cancer followed by colonoscopy at three years and 3-5 years thereafter to detect metachronous cancer. 4. Following adequate clearance of neoplastic polyp(s) survey at 3-5 year intervals. 5. Patients with significant family history of a. Hereditary non polyposis colorectal cancer: colonoscopy every two years beginning at the earlier of age 25, or five years younger than the earliest age of diagnosis of colorectal cancer. Annual colonoscopy should begin at age 40. b. Sporadic colorectal cancer before the age of 60: colonoscopy every five years beginning at age 10 years earlier than the affected relative or every three years if adenoma is found. 6. In patients with ulcerative or Crohn's pancolitis eight or more years' duration or left sided colitis 15 or more years' duration every 1-2 years with systematic biopsies to detect dysplasia.
5. Chronic inflammatory bowel disease of the colon if more precise diagnosis or determination of the extent of activity of disease will influence immediate management.
6. Clinically significant diarrhea of unexplained origin.
7. Intraoperative identification of a lesion not apparent at surgery (e.g., polypectomy site, location of a bleeding site).
8. Treatment of bleeding from such lesions as vascular malformation, ulceration, neoplasia, and polypectomy site (e.g., electrocoagulation, heater probe, laser or injection therapy).
9. Foreign body removal.
10. Excision of colonic polyp.
11. Decompression of acute nontoxic megacolon or sigmoid volvulus.
12. Balloon dilation of stenotic lesions (e.g., anastomotic strictures).
13. Palliative treatment of stenosing or bleeding neoplasms (e.g., laser, electrocoagulation, stenting).
14. Marking a neoplasm for localization.

7.3: GRADE Profiles

Guideline recommendation	Prevalence inappropriate	Evidence supporting recommendation	Study Design	Outcome	Risk of Bias	Indirectness	Inconsistency	Imprecision	Overall
No MRI for non-red flag low back pain	0.1758	Sorby W 1989[a]	Diagnostic accuracy study	Sensitivity	Serious	Very serious	Not serious	Very serious	Very low
				Specificity	Serious	Very serious	Not serious	Very serious	
No CT for non-red flag low back pain	3.69	Wiesel 1986[b]	Prevalence Cohort study	Prevalence of findings in asymptomatic population	Serious	Serious	Not serious	Not serious	Low
No x-rays for non-red flag low back pain	20.22	Kerry 2002[c]	RCT	Roland Morris Disability Scale	Not serious	Not serious	Not serious	Not serious	High
		Kendrick 2001[d]	RCT						
No x-rays for non-red flag low back pain	24	Kerry 2002[c]	RCT	Roland Morris Disability Scale	Not serious	Not serious	Not serious	Not serious	High
		Kendrick 2001[d]	RCT						
No imaging for non-red flag neck pain	22.8	Heller 1983[e]	Prevalence Cohort study	Prevalence of findings in asymptomatic population	Not serious	Serious	Serious	Serious	Very low
		Johnson 1997[f]	Prevalence Cohort study						
No imaging for non-red flag low back pain	23.92	Kerry 2002[c]	RCT	Roland Morris Disability Scale	Not serious	Not serious	Not serious	Not serious	High
		Kendrick 2001[d]	RCT						
No x-rays for	11.79	Kerry 2002[c]	RCT	Roland	Not	Not serious	Not serious	Not	High

non-red flag low back pain		Kendrick 2001[d]	RCT	Morris Disability Scale	serious			serious	
No x-rays for non-red flag low back pain	12	Kerry 2002[c]	RCT	Roland Morris Disability Scale	Not serious	Not serious	Not serious	Not serious	High
		Kendrick 2001[d]	RCT						
No x-rays for non-red flag low back pain	12.27	Kerry 2002[c]	RCT	Roland Morris Disability Scale	Not serious	Not serious	Not serious	Not serious	High
		Kendrick 2001[d]	RCT						
No x-rays for non-red flag low back pain	12.83	Kerry 2002[c]	RCT	Roland Morris Disability Scale	Not serious	Not serious	Not serious	Not serious	High
		Kendrick 2001[d]	RCT						
No x-rays for non-red flag low back pain	12.91	Kerry 2002[c]	RCT	Roland Morris Disability Scale	Not serious	Not serious	Not serious	Not serious	High
		Kendrick 2001[d]	RCT						
No x-rays for non-red flag low back pain	13.01	Kerry 2002[c]	RCT	Roland Morris Disability Scale	Not serious	Not serious	Not serious	Not serious	High
		Kendrick 2001[d]	RCT						
No MRI for non-red flag low back pain	22	Gilbert 2005[g]	RCT	Pain score (Aberdeen and short form 36)	Not serious	Serious	Not serious	Not serious	High
Don't do strept testing if Centor Criteria is 0, 1 or 2	91.54	Centor 1981[h]	Diagnostic accuracy	Sensitivity	Serious	Serious	Serious	Not serious	Very low
				Specificity	Serious	Serious	Serious	Not serious	
		Dagnelie 1998[i]	Diagnostic accuracy	Sensitivity	Serious	Serious	Serious	Not serious	
				Specificity	Serious	Serious	Serious	Not serious	
Brain Natriuretic Peptide (BNP)	95.7	Mueller 2004[J]	RCT	Mortality	Not serious	Serious	Not serious	Serious	Low

to confirm or refute the diagnosis of heart failure									
H.Pylori test to investigate dyspepsia	34.4	Cutler 1995[k]	Diagnostic Accuracy Study	Sensitivity	Very serious	Serious	Not serious	Not serious	Very low
				Specificity	Very serious	Serious	Not serious	Not serious	
In patients with dyspepsia, test and treat for H. pylori before endoscopy	45.12	Bytzer 1994[L]	RCT	Dyspepsia symptoms	Serious	Not serious	Not serious	Serious	Low
		Lewin van den Broek 2001[m]	RCT						
Pulmonary Function Tests to confirm or refute the diagnosis of COPD	25.63	Jackson 2003[n]	Diagnostic accuracy study	Sensitivity	Not serious	Serious	Not serious	Not serious	Low
				Specificity	Not serious	Serious	Not serious	Not serious	
Pulmonary Function Tests to confirm or refute the diagnosis of COPD	55.01	Jackson 2003[n]	Diagnostic accuracy study	Sensitivity	Not serious	Serious	Not serious	Not serious	Low
				Specificity	Not serious	Serious	Not serious	Not serious	
Pulmonary Function Tests to confirm or refute the diagnosis of COPD	73.04	Jackson 2003[n]	Diagnostic accuracy study	Sensitivity	Not serious	Serious	Not serious	Not serious	Low
				Specificity	Not serious	Serious	Not serious	Not serious	
For patients >50 years of age, order endoscopy for dyspepsia	81.15	Talley 1998[o]	Narrative review (Diagnostic accuracy)	Sensitivity	Very serious	Not serious	Not serious	Not serious	Low
				Specificity	Very serious	Not serious	Not serious	Not serious	

Pulmonary Function Tests to confirm or refute the diagnosis of Asthma	76.7	Li 1996[p]	Narrative review (Diagnostic accuracy)	Sensitivity	Very serious	Not serious	Serious	Serious	Very low
				Specificity	Very serious	Not serious	Serious	Serious	
Do not order CT scans for patients with uncomplicated Acute Sinusitis	0.6	Lau 1999[q]	Systematic review (Diagnostic accuracy)	Sensitivity	Not serious	Serious	Serious	Serious	Very low
				Specificity	Not serious	Serious	Serious	Serious	
Order mid-luteal progesterone for workup of infertility	59.4	Wathen 1984[r]	Diagnostic Accuracy Study	Sensitivity	Serious	Serious	Not serious	Serious	Very low
				Specificity	Serious	Serious	Not serious	Serious	
Order semen analysis for workup of infertility	96.9	Opsahl 1996[s]	Diagnostic Accuracy Study	Sensitivity	Serious	Serious	Not serious	Serious	Very low
				Specificity	Serious	Serious	Not serious	Serious	

7.4 References for Appendix table 7.2 and appendix table 7.3

Full references for Guidelines:

1. Brooks P, Bogduk N, Bellamy N, Al E. Evidence-Based Management of Acute Musculoskeletal Pain [Internet]. Australian National Health and Medical Research Council Guideleines. 2003.
2. Brooks P, Bogduk N, Bellamy N, Al E. Evidence-Based Management of Acute Musculoskeletal Pain: A guide for clinicians. Vol. 66, Australian National Health and Medical Research Council Guideleines. 2003.
3. Chou R, Qaseem A, Snow V, Casey D, Cross JT, Shekelle P, et al. Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. Ann Intern Med. 2007 Oct 2;147(7):478–91.
4. Loder E, Weizenbaum E, Frishberg B, Silberstein S. Choosing Wisely in headache medicine: The American Headache Society's list of five things physicians and patients should question. Headache. 2013;53(10):1651–9.
5. Langer-Gould AM, Anderson WE, Armstrong MJ, Langer- AM, Becker A, Getchius TSD. The American Academy of Neurology's Top Five Choosing Wisely recommendations. Neurology. 2013;81:1004–11.
6. Singer PA, Cooper DS, Daniels GH, Ladenson PW, Greenspan FS, Levy EG, et al. Treatment Guidelines for Patients With Thyroid Nodules and Well-Differentiated Thyroid Cancer. Arch Intern Med. 1996;156:2165–72.
7. Bisno AL, Gerber MA, Gwaltney JM, Kaplan EL, Schwartz RH. Practice Guidelines for the Diagnosis and Management of Group A Streptococcal Pharyngitis. Clin Infect Dis. 2002;4:113–25.
8. Patterson CJS, Gauthier S, Bergman H, Cohen CA, Feightner JW, Feldman H, et al. The recognition, assessment and management of dementing disorders: Conclusions from the Canadian consensus conference on dementia. Cmaj. 1999;160(12 SUPPL.).
9. Hendel RC, Berman DS, Di Carli MF, Heidenreich PA, Henkin RE, Pellikka PA, et al. ACCF/ASNC/ACR/AHA/ASE/SCCT/SCMR/SNM 2009 Appropriate Use Criteria for Cardiac Radionuclide Imaging. J Am Coll Cardiol. 2009;53(23):2201–29.
10. Dickstein K, Cohen-Solal A, Filippatos G, McMurray JJV, Ponikowski P, Poole-Wilson PA, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2008 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM). Eur Heart J. 2008;29(19):2388–442
11. Soll A, Achord J, Bozymski G, Brooks S, Lanza F. Medical treatment of peptic ulcer disease: practice guidelines. JAMA. 1996
12. Shaheen NJ, Weinberg DS, Denberg TD, Chou R, Qaseem A, Shekelle P. Upper endoscopy for gastroesophageal reflux disease: best practice advice from the clinical guidelines committee of the American College of Physicians. Ann Intern Med. 2012;157(11):808–16.
13. National Institute for Health and Care Excellence. National clinical guideline on management of chronic obstructive pulmonary disease in adults in primary and secondary care. NICE Clinical Guideline 12. 2004.
14. Pauwels RA, Buist AS, Calverley PMA, Jenkins CR, Hurd SS. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med. 2001;163:1256–76.
15. Rabe KF, Hurd S, Anzueto A, Barnes PJ, Buist SA, Calverley P, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2007;176(6):532–55.

16. American Gastroenterological Association. American Gastroenterological Association Medical Position Statement: Evaluation of Dyspepsia. Vol. 114, Gastroenterology. 1998.
17. U.S. National Institutes of Health. Guidelines for the Diagnosis and Management of Asthma. Vol. 120, National Asthma Education and Prevention Program. 2007.
18. Waldemar G, Dubois B, Emre M, Georges J, McKeith IG, Rossor M, et al. European Federation of Neurological Societies: Recommendations for the diagnosis and management of Alzheimer's disease and other disorders associated with dementia. Eur J Neurol. 2007;14(1):1–26.
19. Canadian Task Force on the Periodic Health Examination. The Canadian Guide to Clinical Preventive Health Care. Health Canada. 1994.
20. American Academy of Otolaryngology. Clinical practice guideline on adult sinusitis. Otolaryngol - Head Neck Surg. 2007;137(3):365–77.
21. Talley NJ, Vakil N. Guidelines for the management of dyspepsia. Am J Gastroenterol. 2005;100(10):2324–37.
22. American Society for Gastrointestinal Endoscopy. Appropriate Use of Gastrointestinal Endoscopy. Gastrointest Endosc. 2000;52(6):831–7.
23. Rubin GP, Meineche-Schmidt V, Roberts AP, Childs SM, de Wit NJ. The European Society of Primary Care Gastroenterology: Management of H. pylori infections. Eur J Gen Pract [Internet]. 1999;5(3):98–104. Available from: http://www.tandfonline.com/doi/full/10.3109/13814789909094271
24. Royal College of Obstetricians and Gynaecologists. The Initial Investigation and Management of the Infertile Couple. 1998
25. Genitourinary Medicine Clinical Effectiveness Group. British Association for Sexual Health and HIV: Management of Epididymitis. 2001.
26. National Institure for Health and Care Excellence (NICE). Urinary Incontinence. NICE Guideline. 2006.

Full references for Evidence:

a) Sorby W. An evaluation of magnetic resonance imaging at The Royal North Shore Hospital of Sydney, 1986-1987. Med J Aust. 1989;151(1):8–11.
b) Wiesel SW (1986). A study of computer-assisted tomography. The incidence of positive CAT scans in an asymptomatic group of patients. Spine, 9: 549–551 (6495024)
c) Kerry S, Hilton S, Dundas D, Rink E, Oakeshott P (2002). Radiography for low back pain: a randomised controlled trial and observational study in primary care. The British Journal of General Practice, 52: 469–474. (12051211)
d) Kendrick D, Fielding K, Bentley E, Kerslake R, Miller P, Pringle M (2001). Radiography of the lumbar spine in primary care patients with low back pain: randomised controlled trial. BMJ, 322: 400–405. (11179160)
e) Heller CA, Stanley P, Lewis-Jones B, Heller RF (1983). Value of x-ray examinations of the cervical spine. BMJ, 287: 1276–1278.
f) Johnson MJ, Lucas GL (1997). Value of cervical spine radiographs as a screening tool. Clinical Orthopaedics and Related Research, 340: 102–108.
g) Gilbert FJ, Grant AM, Gillan MGC, Vale LD, Campbell MK, Scott NW, et al. Low Back Pain: Influence of Early MR Imaging or CT on Treatment and Outcome—Multicenter Randomized Trial. Radiology. 2004;231(2):343–51.
h) Centor RM, Witherspoon JM, Dalton HP, Brody CE, Link K. The diagnosis of strep throat in adults in the emergency room. Med Decis Making. 1981;1(3):239–46.
i) Dagnelie CF, Bartelink ML, Van Der Graaf Y, Goessens W, De Melker RA. Towards a better diagnosis of throat infections (with group A β-haemolytic streptococcus) in general practice. Br J Gen Pract. 1998;48(427):959–62.
j) Mueller C, Laule-Kilian K, Scholer A, Frana B, Rodriguez D, Schindler C, et al. Use of B-

type natriuretic peptide for the management of women with dyspnea. Am J Cardiol. 2004;94(12):1510–4.
k) Cutler AF, Havstad S, Ma CK, Blaser MJ, Perez-Perez GI, Schubert TT. Accuracy of invasive and noninvasive tests to diagnose Helicobacter pylori infection. Gastroenterology. 1995;109(1):136–41.
l) Bytzer P, Hansen JM o ller, Schaffalitzky de Muckadell OB. Empirical H2-blocker therapy or prompt endoscopy in management of dyspepsia. Lancet. 1994;343(8901):811–6.
m) Lewin-Van Den Broek NT, Numans ME, Buskens E, Verheij TJM, De Wit NJ, Smout AJPM. A randomised controlled trial of four management strategies for dyspepsia: Relationships between symptom subgroups and strategy outcome. Br J Gen Pract. 2001;51(469):619–24.
n) Jackson H, Hubbard R. Detecting chronic obstructive pulmonary disease using peak flow rate: cross sectional survey. BMJ. 2003;327(7416):653–4
o) Talley NJ, Silverstein MD, Agréus L, Nyrén O, Sonnenberg A, Holtmann G, et al. AGA technical review: Evaluation of dyspepsia. Gastroenterology. 1998;114(3):582–95
p) Li JTC, O'Connell EJ. Clinical evaluation of asthma. Ann Allergy, Asthma, Immunol. 1996;76(1):1–14.
q) Lau J, Zucker D, Engels EA, et al. Diagnosis and Treatment of Acute Bacterial Rhinosinusitis. Rockville (MD): Agency for Health Care Policy and Research (US); 1999 Mar. (Evidence Reports/Technology Assessments, No. 9.)
r) Wathen NC, Perry L, Lilford RJ, Chard T. Interpretation of single progesterone measurement in diagnosis of anovulation and defective luteal phase: observations on analysis of the normal range. Br Med J (Clin Res Ed). 1984 Jan;288(6410):7–9.
s) Opsahl MS, Dixon NG, Robins ER, Cunningham DS. Single vs. multiple semen specimens in screening for male infertility factors. A comparison. J Reprod Med. 1996 May;41(5):313–5.

7.5: 23 questions in AGREE II tool

Domain 1. 'Scope and Purpose' is concerned with the overall aim of the guideline, the specific health questions, and the target population (items 1-3).

1. The overall objective(s) of the guideline is (are) specifically described (1-7)
2. The health question(s) covered by the guideline is (are) specifically described (1-7)
3. The population (patients, public, etc.) to whom the guideline is meant to apply is specifically described (1-7)

Domain 2. 'Stakeholder Involvement' focuses on the extent to which the guideline was developed by the appropriate stakeholders and represents the views of its intended users (items 4-6).

4. The guideline development group includes individuals from all relevant professional groups (1-7)
5. The views and preferences of the target population (patients, public, etc.) have been sought (1-7)
6. The target users of the guideline are clearly defined (1-7)

Domain 3. 'Rigour of Development' relates to the process used to gather and synthesize the evidence, the methods to formulate the recommendations, and to update them (items 7-14).

7. Systematic methods were used to search for evidence (1-7)
8. The criteria for selecting the evidence are clearly described (1-7)
9. The strengths and limitations of the body of evidence are clearly described (1-7)
10. Methods for formulating the recommendations are clearly described (1-7)
11. The health benefits, side effects, and risks have been considered in formulating the recommendations (1-7)
12. There is an explicit link between the recommendations and the supporting evidence (1-7)
13. The guideline has been externally reviewed by experts prior to its publication (1-7)
14. A procedure for updating the guideline is provided (1-7)

Domain 4. 'Clarity of Presentation' deals with the language, structure, and format of the guideline (items 15-17).

15. The recommendations are specific and unambiguous (1-7)
16. The different options for management of the condition or health issue are clearly presented (1-7)
17. Key recommendations are easily identifiable (1-7)

Domain 5. 'Applicability' pertains to the likely barriers and facilitators to implementation, strategies to improve uptake, and resource implications of applying the guideline (items 18-21).

18. The guideline describes facilitators and barriers to its application
19. The guideline provides advice and/or tools on how the recommendations can be put into practice
20. The potential resource implications of applying the recommendations have been considered
21. The guideline presents monitoring and/or auditing criteria

Domain 6. 'Editorial Independence' is concerned with the formulation of recommendations not being unduly biased with competing interests (items 22-23).

22. The views of the funding body have not influenced the content of the guideline
23. Competing interests of guideline development group members have been recorded and addressed

Overall assessment

Rate overall quality of guideline (1-7)

7.6 AGREE II Formula

Formula to calculate score for each Domain:

$$\frac{\text{Obtained score} - \text{Minimum possible score}}{\text{Maximum possible score} - \text{Minimum possible score}}$$

Where:

- Obtained score = domain score for appraiser 1 + domain score for appraiser 2. *Domain score is the sum of each question.
- Maximum possible score = 7 (strongly agree) x number of items in individual domain x number of appraisers
- Minimum possible score = 1 (strongly disagree) x number of items in individual domain x number of appraisers

For example, Domain 1: Scope and Purpose

	Question 1: The overall objective(s) of the guideline is (are) specifically described (1-7)	Question 2: The health question(s) covered by the guideline is (are) specifically described (1-7)	Question 3: The population (patients, public, etc.) to whom the guideline is meant to apply is specifically described (1-7)	Total Domain Score
Appraiser 1	5	6	6	17
Appraiser 2	6	6	7	19
Obtained score				36

$$\frac{36 - (1x3x2)}{(7x3x2) - (1x3x2)}$$

$$\therefore$$

$$\frac{36 - 6}{42 - 6}$$

$$\therefore$$

$$\frac{30}{36}$$

$$= 83\%$$

Appendix Chapter 8

8.1 Rank of 44 tests by temporal increase

Test	Rank by temporal increase
Vitamin D	1
Knee MRI	2
Pelvic CT	3
Brain MRI	4
Urine Albumin Creatinine ratio	5
Iron	6
Ferritin	7
Folate	8
Vitamin B12	9
C-Reactive Protein	10
Lumbar Spine MRI	11
Testosterone	12
HbA1c	13
Clotting	14
DEXA	15
Bone Profile	16
Urine Non-illicit Drug monitoring	17
Knee x-ray	18
Liver Function Tests	19
Pelvic Ultrasound	20
Full Blood Count	21
Renal Function	22

Prostate Specific Antigen	23
Progesterone	24
Chest x-ray	25
Brain CT	26
Female Sex Hormone	27
Upper Endoscopy	28
Thyroid Function Tests	29
Urine MCS	30
Colonoscopy	31
Erythrocyte sedimentation rate	32
Echocardiogram	33
Oestradiol	34
Troponin	35
Urine Dipstick	36
Lipids	37
Lumbar spine x-ray	38
Glucose	39
Urine Albumin	40
Spirometry	41
Pap Smear	42
Vaginal Swab	43
Creatinine Kinase	44

8.2 Rank of 44 tests by frequency of use

Test	Rank by frequency of use
Renal Function	1
Full Blood Count	2
Liver Function Tests	3
Lipids	4
Thyroid Function Tests	5
Glucose	6
HbA1c	7
Clotting	8
Urine Dipstick	9
Bone Profile	10
Erythrocyte sedimentation rate	11
Ferritin	12
C-Reactive Protein	13
Urine MCS	14
Vitamin B12	15
Folate	16
Spirometry	17
Urine Albumin Creatinine ratio	18
Chest x-ray	19
Prostate Specific Antigen	20
Vitamin D	21
Urine Albumin	22
Iron	23
Female Sex Hormone	24

Pelvic Ultrasound	25
Knee x-ray	26
Creatinine Kinase	27
Upper Endoscopy	28
Echocardiogram	29
Colonoscopy	30
Pap Smear	31
Testosterone	32
Lumbar Spine MRI	33
Lumbar spine x-ray	34
Vaginal Swab	35
Oestradiol	36
DEXA	37
Progesterone	38
Brain CT	39
Troponin	40
Urine Non-illicit Drug monitoring	41
Brain MRI	42
Knee MRI	43
Pelvic CT	44

8.3 Rank of 44 tests by between-practice variation

Test	Rank by CoV (Most to least)
Non-illicit Drug monitoring	1
Urine Albumin	2
Pelvic CT	3
Pap Smear	4
Troponin	5
Creatinine Kinase	6
Brain MRI	7
Vitamin D	8
Clotting	9
Prostate Specific Antigen	10
Vaginal Swab	11
DEXA	12
Knee MRI	13
Lumbar Spine MRI	14
Brain CT	15
Colonoscopy	16
Bone Profile	17
Progesterone	18
Upper Endoscopy	19
Iron	20
C-Reactive Protein	21
Urine MCS	22
Lumbar spine x-ray	23
Chest x-ray	24

Urine Albumin Creatinine ratio	25
Knee x-ray	26
Oestradiol	27
Pelvic Ultrasound	28
Echocardiogram	29
Vitamin B12	30
Folate	31
Urine Dipstick	32
Liver Function Tests	33
Erythrocyte sedimentation rate	34
Renal Function	35
Full Blood Count	36
Glucose	37
HbA1c	38
Lipids	39
Ferritin	40
Spirometry	41
Thyroid Function Tests	42
Female Sex Hormone	43
Testosterone	44

8.4 Combined rank (Temporal, between-practice variation and frequency of use) of all 44 tests

Test	Rank
Vitamin D	1
Clotting	2
Vitamin B12	3
Bone Profile	4
C-Reactive Protein	5
Urine Albumin Creatinine ratio	6
Iron	7
Pelvic CT	8
Brain MRI	9
Prostate Specific Antigen	10
Folate	11
Liver Function Tests	12
HbA1c	13
Knee MRI	14
Lumbar Spine MRI	15
Renal Function	16
Ferritin	17
Full Blood Count	18
Non-illicit Drug monitoring	19
DEXA	20
Urine Albumin	21
Urine MCS	22
Chest x-ray	23

Knee x-ray	24
Pelvic Ultrasound	25
Upper Endoscopy	26
Thyroid Function Tests	27
Colonoscopy	28
Creatinine Kinase	29
Erythrocyte sedimentation rate	30
Pap Smear	31
Urine Dipstick	32
Brain CT	33
Lipids	34
Progesterone	35
Troponin	36
Glucose	37
Testosterone	38
Echocardiogram	39
Female Sex Hormone	40
Lumbar spine x-ray	41
Oestradiol	42
Spirometry	43
Vaginal Swab	44